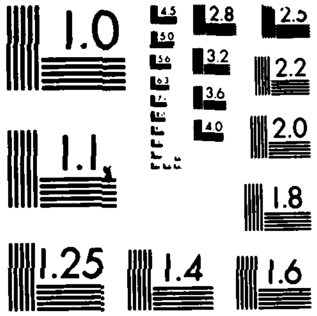
AD-A142 255   SENSITIVITY BASED SEGMENTATION AND IDENTIFICATION IN   1/1
AUTOMATIC SPEECH REC..(U) VERMONT UNIV BURLINGTON DEPT
OF COMPUTER SCIENCE AND ELECTRIC.. R ABSHER 30 MAR 84
UNCLASSIFIED   AFOSR-TR-84-0476 AFOSR-83-0053   F/G 17/2   NL

END
DTIC
7-84



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

The University of Vermont
COLLEGE OF ENGINEERING AND MATHEMATICS
COMPUTER SCIENCE AND ELECTRICAL ENGINEERING DEPARTMENT
VOTEY BUILDING
BURLINGTON, VERMONT 05405-0156
(802) 656-3330

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

FINAL REPORT

SENSITIVITY BASED SEGMENTATION AND IDENTIFICATION IN AUTOMATIC SPEECH RECOGNITION

ORGANIZATION:

The University of Vermont
85 South Prospect Street
Burlington, VT 05405

PRINCIPAL INVESTIGATOR:

Richard Absher, Ph.D.
Computer Science and Electrical Engineering Department

GRANT INFORMATION:

Grant No. AFOSR - 83 - 0053
Purchase Request No. FQ 8671 - 8300537
Project Task 2304/D9
Grant Period - January 1, 1983 through July 31, 1983

DATE:

March 30, 1984

ADA142255

SENSITIVITY BASED SEGMENTATION AND IDENTIFICATION IN AUTOMATIC SPEECH RECOGNITION

Dr. Richard G. Absher
Principal Investigator

ABSTRACT

This research program continued an investigation of sensitivity analysis, and its use in the segmentation and identification of the phonetic units of speech, that was initiated during the 1982 Summer Faculty Research Program. The elements of the sensitivity matrix, which express the relative change in each pole of the speech model to a relative change in each coefficient of the characteristic equation, were evaluated for an expanded set of data which consisted of six vowels contained in single words spoken in a simple carrier phrase by five males with differing dialects. The objectives were to evaluate the sensitivity matrix, interpret its changes during the production of the vowels, and to evaluate inter-speaker variations. It was determined that the sensitivity analysis (1) serves to segment the vowel interval, (2) provides a measure of when a vowel is "on target," and (3) should provide sufficient information to identify each particular vowel. Based on the results presented, sensitivity analysis should result in more accurate segmentation and identification of phonemes and should provide a practicable framework for incorporation of acoustic-phonetic variance as well as time and talker normalization.

TABLE OF CONTENTS

DESCRIPTION OF RESEARCH PROGRAM

I. INTRODUCTION:

There are several general approaches used in current computer based continuous speech-recognition systems. A typical system uses a three-step procedure.[33] In the first step, a word or phrase is divided into time segments or frames. For each frame, a "best fit" is determined for a particular parametric representation (speech model). Next, a statistical decision rule is used to tentatively determine (or estimate) the phoneme corresponding to each frame. Third, a set of phonological decision rules is used to combine the phonemic decisions of the several frames and access lexical candidates. As a contrast, Klatt[12] proposed a system in which samples of the speech waveform are analyzed to determine a sequence of spectral representations which are directly decoded into lexical candidates by a network constructed from phonemic, phonetic, and phonological rules.

Regardless of the speech processing system used, Klatt[12] has described eight problem areas that must be overcome. These are (1) acoustic-phonetic variance, (2) segmentation of the signal into phonetic units, (3) time normalization, (4) talker normalization, (5) lexical representations for optimal search, (6) phonological recoding of words in sentences, (7) dealing with errors in the initial phonetic representation during lexical matching, and (8) interpretation of prosodic cues to lexical items and sentence structure.

This research program was focused on problem 2 and its interplay with problems 1, 3, and 4. That is, this research project investigated a novel 2-level scheme for more accurate segmentation of the speech signal into phonetic units. It was anticipated that this scheme would allow for a more accurate application of decision rules, and would provide a practicable framework for incorporation of acoustic-phonetic variance as well as time and talker normalization. Although the sensitivity analysis is expressed within the framework of a three-step speech analysis system, it could also be viewed as an alternative to the sequence of spectra as utilized by Klatt. Furthermore, the techniques developed for continuous speech-recognition systems may also provide important implementation advantages when compared with current isolated word or word spotting systems.

Since the current study was limited to non-nasal vowels, the initial parametric representation of each frame consisted of the linear prediction coefficients which were used to express the coefficients of the following characteristic equation:

$$q(s) = s^n + a_2 s^{n-1} + \dots a_n s + a_{n+1} \qquad (1)$$

The second level of the segmentation depended on the evaluation and interpretation of a sensitivity matrix with elements defined as:

$$S(k,i) = \frac{a_k}{r_i} \frac{dr_i}{da_k} \qquad (2)$$

This definition easily leads to the following closed-form expression:

$$S(k,i) = \frac{1}{k-1-n + \sum_{j=1}^{n} \frac{r_i}{r_i + P_{jk}}} \qquad \begin{array}{l} i = 1, \ldots, n \\ k = 2, \ldots, n + 1 \end{array} \qquad (3)$$

where r_i is a root of the characteristic equation and P_{jk} is the j-th root of the characteristic equation when its k-th coefficient is assigned the value zero. Thus S(k,i) expresses the relative change in the location of filter pole i to the relative change in coefficient a_k.

II. SPECIFIC OBJECTIVES:

This research project focused on a selected set of six vowels contained in single words, spoken in a simple carrier phrase, by five males with differing dialects. The specific objectives were to evaluate and interpret the changes in the sensitivity matrix that occur during the production of the vowels, to use this broader data set to test the conclusions reached during the Summer Faculty Research period, and to evaluate inter-speaker variations.

It was necessary to (1) record, digitize, and frame the data; (2) calculate the coefficients of the characteristic equation for each frame via linear predictive analysis; (3) evaluate the sensitivity matrix for each frame; (4) determine if the sensitivity analysis provided a measure of the degree to which a vowel was "on target"; (5) determine if the sensitivity matrix can be used to identify the individual phonemes; and (6) evaluate inter-speaker variations.

III. GENERAL BACKGROUND:

Automatic speech recognition (ASR) has potential application to various USAF operational problems. Examples include voice control of devices and systems, intelligence data handling, and language identification. A practical ASR system "must operate on the continuous utterance of any number of speakers in moderate or even poor noise environments."[1] Major sources of difficulty include acoustic-phonetic variance and segmentation of the acoustic signal. The following discussion reviews the reasons for this difficulty and expresses the research problem.

Connected-speech-recognition systems have utilized techniques that represent sound patterns in smaller linguistic units than words; one being in terms of phonemes.[13] However, the location and specification of the acoustic characteristics of phonemes has been a central problem to speech recognition.[2] The problem is not attributable to mechanical limitations. Instead, it is considered to be the result of the very nature of human speech production and perception, for phonemes are not individual sounds, but rather classes of acoustically different sounds which speakers of a language have learned to call equivalent.[3]

In contrast to the discrete units of linguistic analysis (i.e., sentences, phrases, words, morphemes, phonemes), the acoustic representation of an utterance is semi-continuous. The problem for those studying speech recognition is to map semi-continuous acoustic waveforms into discrete linguistic units.[4] Attempts to accomplish this task have revealed a number of sources of variation that make this mapping difficult. Lack of a one-to-one correspondence between acoustic segments and the linguistic units they represent can be attributable to (1) coarticulation, (2) allophonic variation, (3) stress and rate of speech production, (4) individual speaker differences, and (5) dialectical variation.

Coarticulation can be defined as the influence of one phoneme upon another.[5] Logically, the mismatch between phoneme and acoustic representation caused by coarticulation can be divided into cases in which (1) a phoneme has more than one acoustic representation, and (2) an acoustic segment can represent more than one phoneme. An example of the first case are vowels that have nasal cavity resonances and antiresonance when produced in nasal consonant contexts (e.g., m<u>a</u>n) but not in others (e.g., p<u>a</u>t).[6,7] Speakers of a language group these different acoustic events into the same phoneme class,[3] and so must speech recognition systems.

An example of the second kind of mismatch,in which a unique acoustic signal can correspond to one of several phonemes, involves the noise burst frequency due to place of articulation of stop consonants in CV (consonant-vowel) syllables. It has been demonstrated that a noise burst of a particular frequency is perceived by a listener as /p/ when followed by /i/. The same noise burst is perceived by a listener as /k/ when the following vowel was /æ/.[8] As a consequence of multiple linguistic interpretations of the same acoustic segment, it is imperative that speech recognition systems be able to defer decisions about the phonemic identity of an acoustic event until context can be considered.

Allophonic variation refers to the language specific systematic use of different sound segments (phones) to represent a particular phoneme. For example, the voiceless stop consonants /p/, /t/, /k/ have three allophones in English which occur in contexts specified by definable rules.[9] Aspirated voiceless stops (produced with an audible puff of air at release) are used at the beginning of stressed syllables (e.g., pea). Unaspirated stops (produced without the audible release of air) are used (a) at the beginning of unstressed syllables (e.g., appear), (b) in clusters with /s/ (e.g., speak), and (c) at the ends of words (e.g., keep). Unreleased stops (produced without opening the vocal tract after closure) are used (a) when the stop consonant precedes a homorganic (same place) consonant (e.g., keep me), and (b) optionally at the end of a phrase (e.g., keep).

Stress and rate of production affect the way in which speech sounds are articulated and, as a consequence, affect their acoustic representations. Under conditions of increased speech rate, the duration of some speech segments is decreased, and the articulatory targets achieved typically "fall short."[10] Stressed segments, on the other hand, are known to be longer and to more closely approximate articulatory targets.[9]

Inter-speaker differences in physical structure are another source of acoustic-linguistic mismatch. The differences in acoustic output that arise from differences in the physical structures themselves are predictable from the acoustic theory of production.[11] In this widely accepted view, the vocal tract is considered to be a resonating tube where movement of the vocal structures alters the shape of the tube and results in the production of different sounds. Important considerations are age and sex of the speaker, since both of these parameters influence the size of the larynx and the size of the vocal tract, causing differences in fundamental frequency and formant frequencies.

Dialectical variations include the use of different phonemic contrasts by speakers of subgroups of a language.[9] These variations are historically derived from different language backgrounds and geographical isolation of populations of speakers of the language. The phonemic differences found are concentrated in the vowels and r-like sounds of English. As such, dialectical variation is an important consideration in any automatic recognition scheme designed to identify vocalic productions.

IV. APPROACH:

General.

The acoustic theory of speech production considers the vocal tract as a resonating tube that filters the sound produced by one of a variety of sources, primarily the many forms of phonation produced by the larynx.[9] For non-nasal vowels, an approximate representation of the filter is:[14]

$$T(s) = \prod_{i=1}^{n/2} \frac{r_i r_i^*}{(s + r_i)(s + r_i^*)} \qquad (4)$$

where the constant r_i and its complex conjugate r_i^* are determined by the values of the i-th formant frequency f_i and its bandwidth bw_i. That is,

$$r_i = \pi bw_i + j2\pi f_i \qquad (5)$$

Kenneth Stevens has used simple acoustic tube models to investigate the interrelationships between the shape of the vocal tract and the various formant frequency and bandwidth changes. As a result, Stevens proposed that there is a quantal nature to speech. That is, "there are certain articulatory conditions for which a small change in some parameter describing the articulation gives rise to an apparently large change in the acoustic characteristics of the output; there are other conditions for which substantial perturbations of certain aspects of the articulation produce negligible changes in the characteristics of the acoustic signal."[18]

For the high front vowel /i/, Stevens' acoustic analysis predicts a low first formant and that formants 2 and 3 should be close together. Furthermore, he concluded that for the low and high back vowels /a, u/, formants 1 and 2 should be close. In the following discussion, the sensitivity matrix is proposed as a method to locate and characterize these kinds of formant interrelations.

Expanding the denominator of T(s) gives the characteristic equation:

$$q(s) = s^n + a_2 s^{n-1} + \ldots + a_n s + a_{n+1} \qquad (6)$$

If any coefficient a_i is varied, then each constant [illegible] must also change. The sensitivity matrix,[15] defined in equation (2), is a relative measure for the extent of these changes. As illustrated in Figure 1 for the case where n equals six, it is possible to vary each coefficient, one at a time, from zero to infinity and make a sketch of the corresponding roots (root-locus) of the characteristic equation.[16] These root changes reflect, as described by equation (5), the changes in formant frequencies and formant bandwidths.

Note that the elements S(i,j) of the sensitivity matrix are proportional to the slopes of root-locus branches at the points corresponding to the particular coefficient values. Thus the elements of the sensitivity matrix are complex quantities which express the magnitude and direction of root changes due to coefficient changes. Because S(i,j) is normalized, a direction or phase of 0 means that the root is moving in the direction of the vector from the s-plane origin to the root.

For example, at the points labeled 1 on the curves for a_2 shown in Figure 1, increasing a_2 results in small changes in the three formant frequencies, but significantly changes the

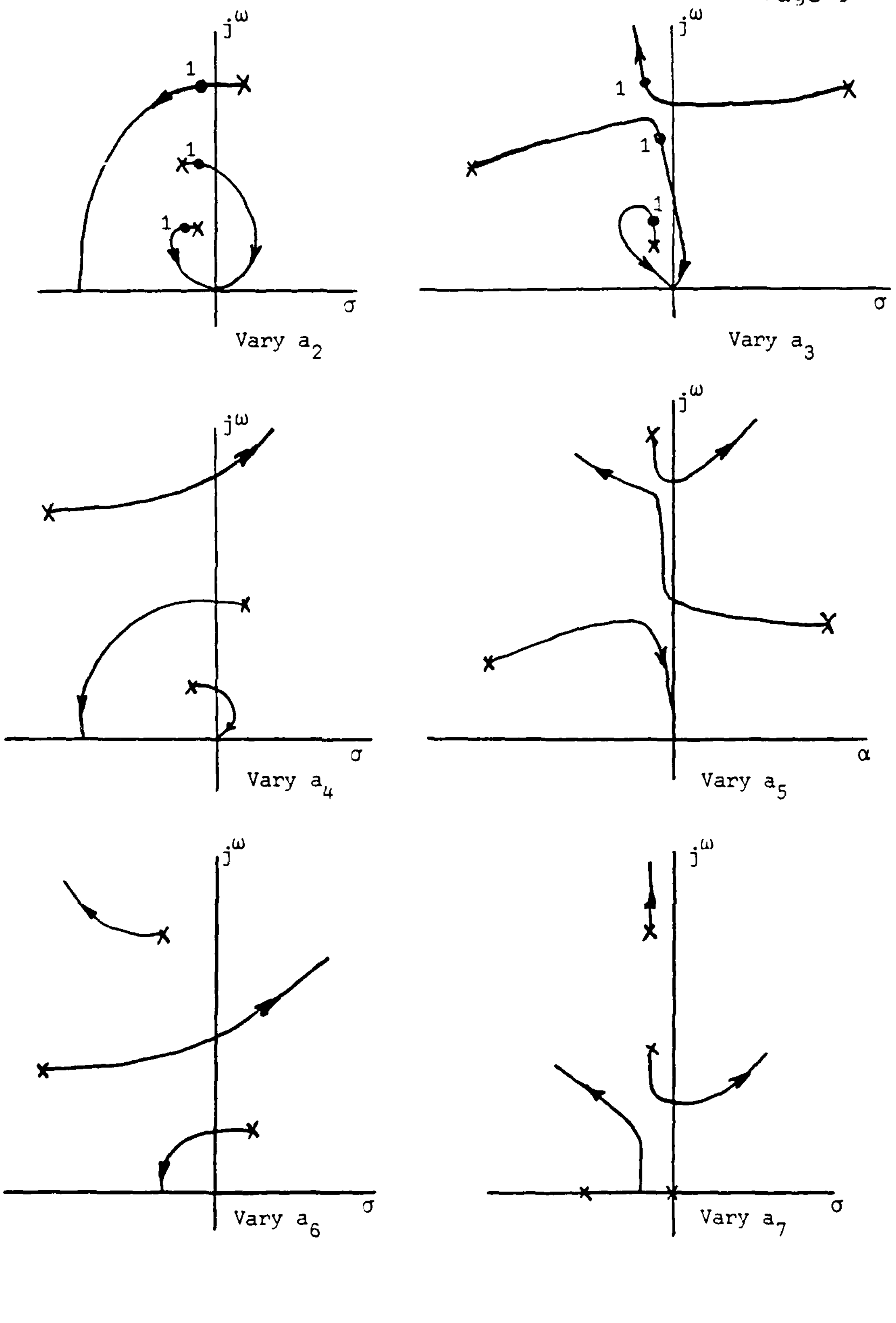


FIGURE 1 - ROOT-LOCUS SKETCHES

bandwidths. Expressed in terms of sensitivity elements, S(2,1) and S(2,3) have a phase of essentially 90 degrees, whereas S(2,2) has a phase of essentially 270 degrees. By contrast, at the points labeled 1 on the curves for a_3 shown in Figure 1, increasing a_3 results in small changes in bandwidth, but significant changes in formant frequencies. Expressed in terms of sensitivity elements, S(3,1) and S(3,3) have a phase of essentially 0 degrees, whereas S(2,2) has a phase of essentially 180 degrees.

Previous Work.

The investigations initiated during the 1982 Summer Faculty Research Program have utilized vowel data reported by Dennis Klatt[14] and shown in Figure 2. These data are for his voice and are a composite obtained from the analysis of many consonant-vowel productions. The parameters listed are the initial and final values of the first three formants and their bandwidths. Each vowel is represented by Klatt with a two letter code which will be used throughout this paper. The correspondence between this code and a more standard phonetic transcription is seen in Figure 2. A computer program was written to

(1) calculate, from the given formant frequencies and bandwidths, the coefficients of the characteristic equation,

(2) vary each coefficient a_i by ±25% or ±50% from its initial or nominal value,

(3) calculate the corresponding elements of the sensitivity matrix and,

(4) make plots for the magnitude and phase of each sensitivity element as each coefficient a_i is varied. All the vowel data of Figure 2 has been processed and the results are briefly summarized in the following discussion. The final report contains a more complete discussion.[17]

D.H. Klatt

Table 2. Parameter values for the synthesis of selected vowels. If two values are given, the vowel is diphthongized or has a schwa-like offglide in the speech of the author. The amplitude of voicing, AV, and fundamental frequency, F0, must also be given contours appropriate for an isolated vowel.

Vowel	F1	F2	F3	B1	B2	B3
IY i	310	2020	2960	45	200	400
	290	2070	2960	60	200	400
IH ɪ	400	1800	2570	50	100	140
	470	1600	2600	50	100	140
EY e	480	1720	2520	70	100	200
	330	2020	2600	55	100	200
EH ɛ	530	1680	2500	60	90	200
	620	1530	2530	60	90	200
AE æ	620	1660	2430	70	150	320
	650	1490	2470	70	100	320
AA ɑ	700	1220	2600	130	70	160
AO ɔ	600	990	2570	90	100	80
	630	1040	2600	90	100	80
AH ʌ	620	1220	2550	80	50	140
OW o	540	1100	2300	80	70	70
	450	900	2300	80	70	70
UH ʊ	450	1100	2350	80	100	80
	500	1180	2390	80	100	80
UW u	350	1250	2200	65	110	140
	320	900	2200	65	110	140
ER ɝ	470	1270	1540	100	60	110
	420	1310	1540	100	60	110
AY aʸ	660	1200	2550	100	70	200
	400	1880	2500	70	100	200
AW aʷ	640	1230	2550	80	70	140
	420	940	2350	80	70	80
OY oʸ	550	960	2400	80	50	130
	360	1820	2450	60	50	160

FIGURE 2 - FORMANT FREQUENCIES AND BANDWIDTHS OF SELECTED VOWELS (reproduced from reference 14, page 291)

Figures 3 and 4 show the magnitude and phase plots for S(2,*) for the high front vowel /IY/ as coefficient a(2) is varied ±25% from its nominal value. In all plots, formants 1, 2, and 3 correspond to plotting symbols *,△, and □. Thus, Figure 3 shows that formant 3 is the most sensitive and formant 1 is the least sensitive to changes in coefficient a(2). But even a sensitivity of .2 is small. Referring to Figure 4, the phase curves of formants 1 and 3 are essentially 90 degrees, and the phase of formant 2 is essentially 270 degrees. Thus, an increase in coefficient a(2) increases the bandwidths of formants 1 and 3, decreases the bandwidth of formant 2, and essentially does not change any of the formant frequencies. This type of influence was found to hold for all vowels and also for coefficients a(4) and a(6).

For the same phoneme /IY/, Figures 5 through 8 show the magnitude and phase plots for sensitivity elements S(3,*) and S(5,*) as coefficients a(3) and a(5) are varied. Figures 6 and 8 show that under the nominal conditions of 1.0 a(3) and 1.0 a(5), the phase associated with each formant is essentially 0 or 180 degrees. A phase of 0 means that the root is moving in the direction of the vector from the s-plane origin to the root.

Referring to Figure 6, formant frequencies 1 and 3 are essentially increasing, formant frequency 2 essentially decreasing, and only small changes are occurring in the formant bandwidths as coefficient a(3) increases. As shown in Figure 8, the same kind of changes occur with a decrease in coefficient a(5). This kind of influence was also found to hold for coefficient a(7) and for all the vowels.

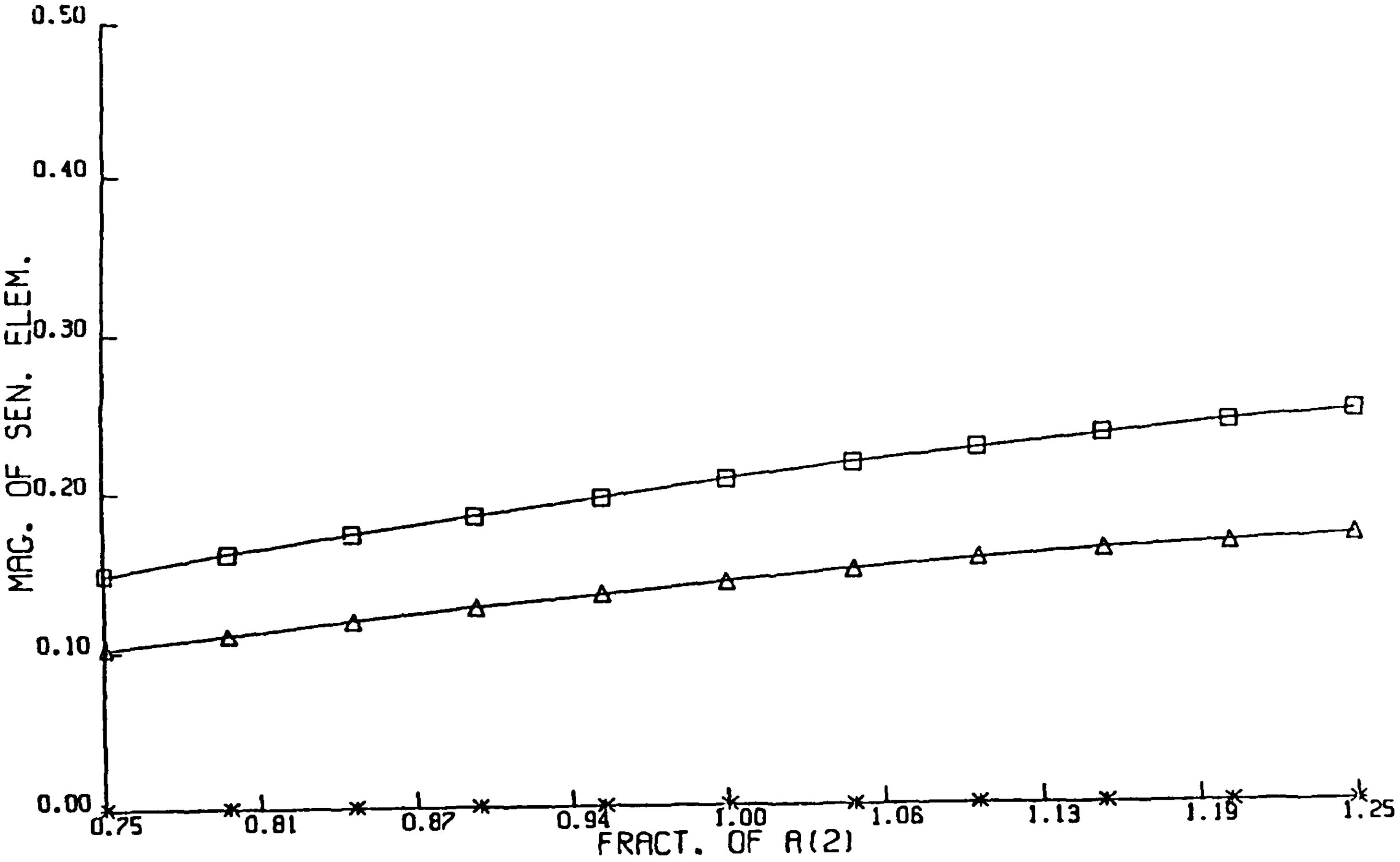


FIGURE 3 - PHONENE IY (1) S(2,*)

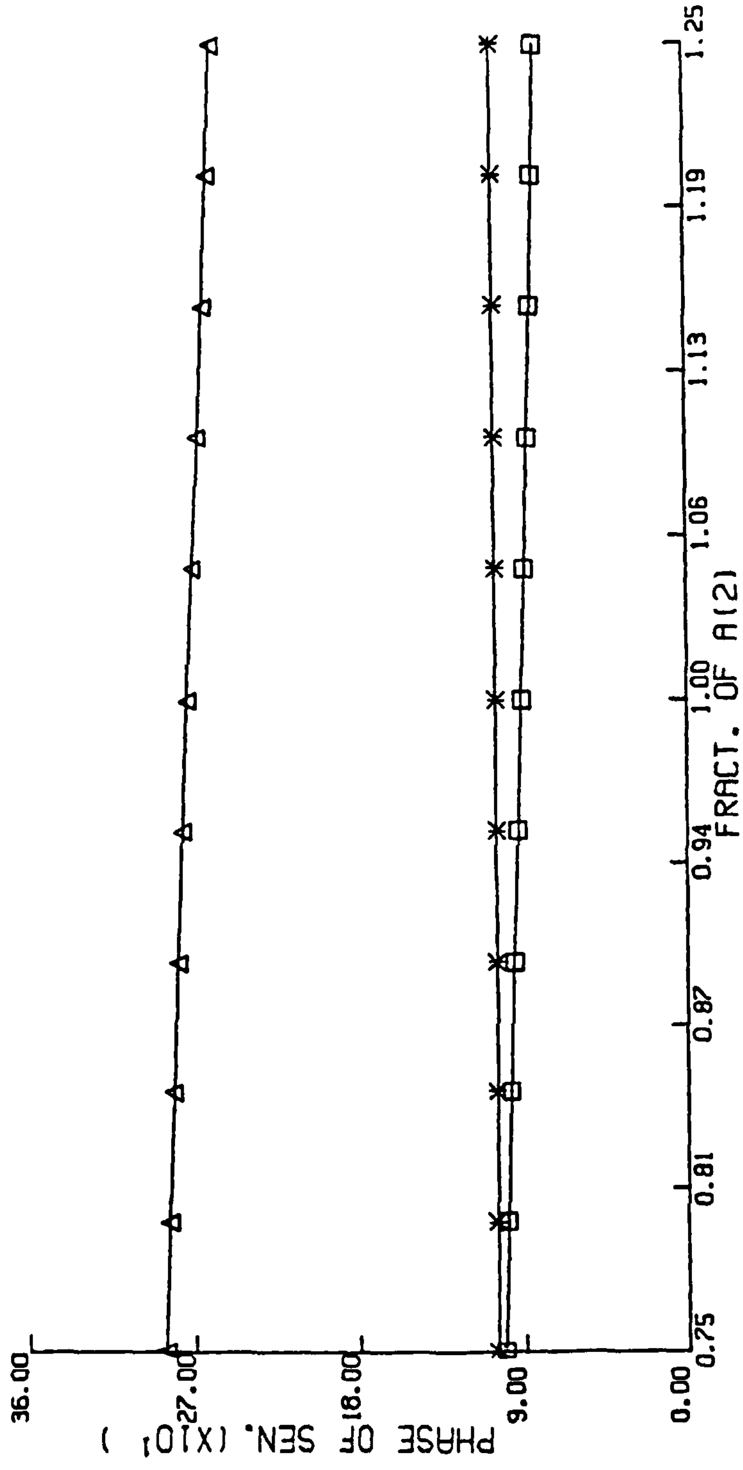


FIGURE 4 - PHONEME IY (1) S(2,*)

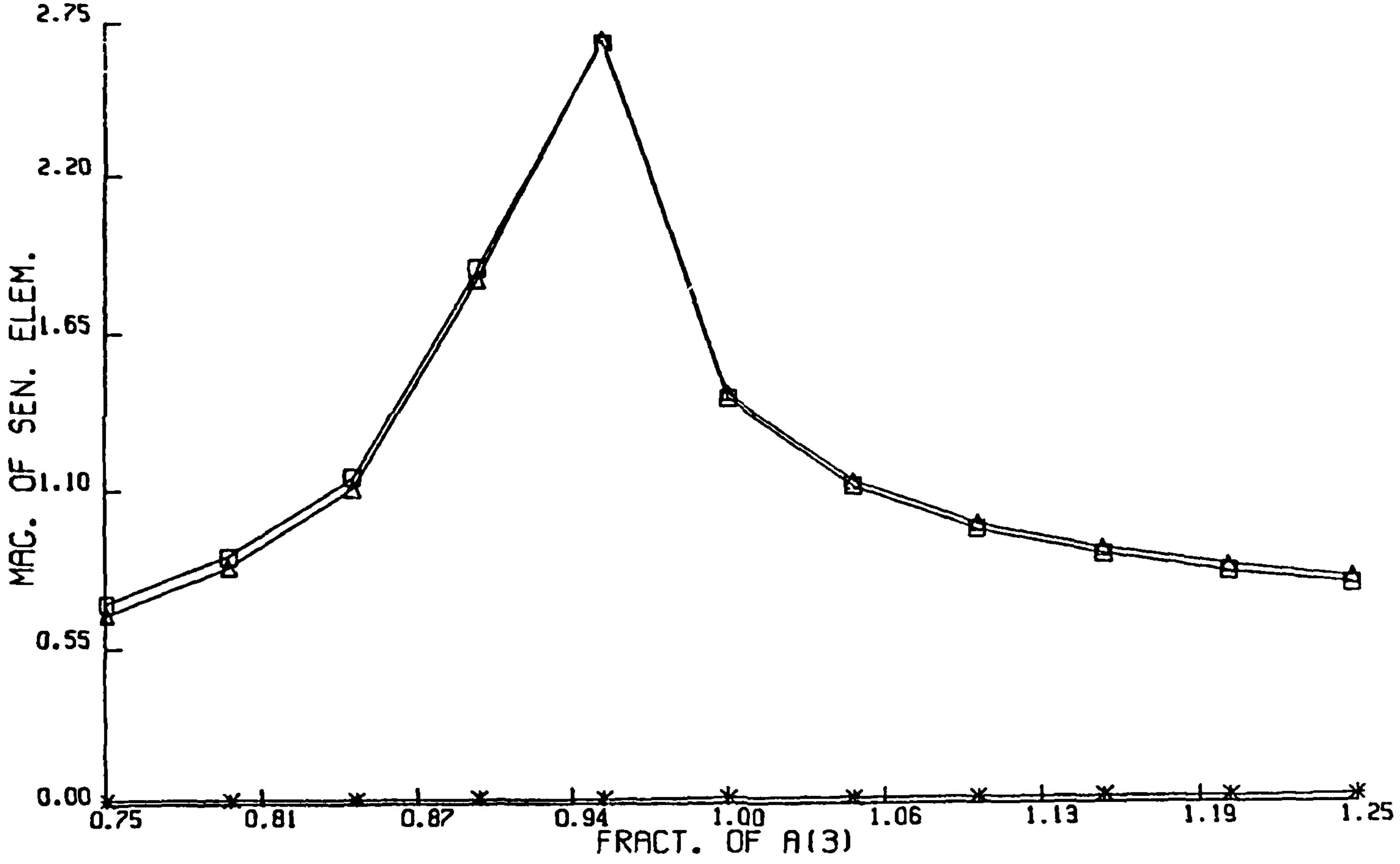


FIGURE 5 - PHONEME IY (1) S(3,*)

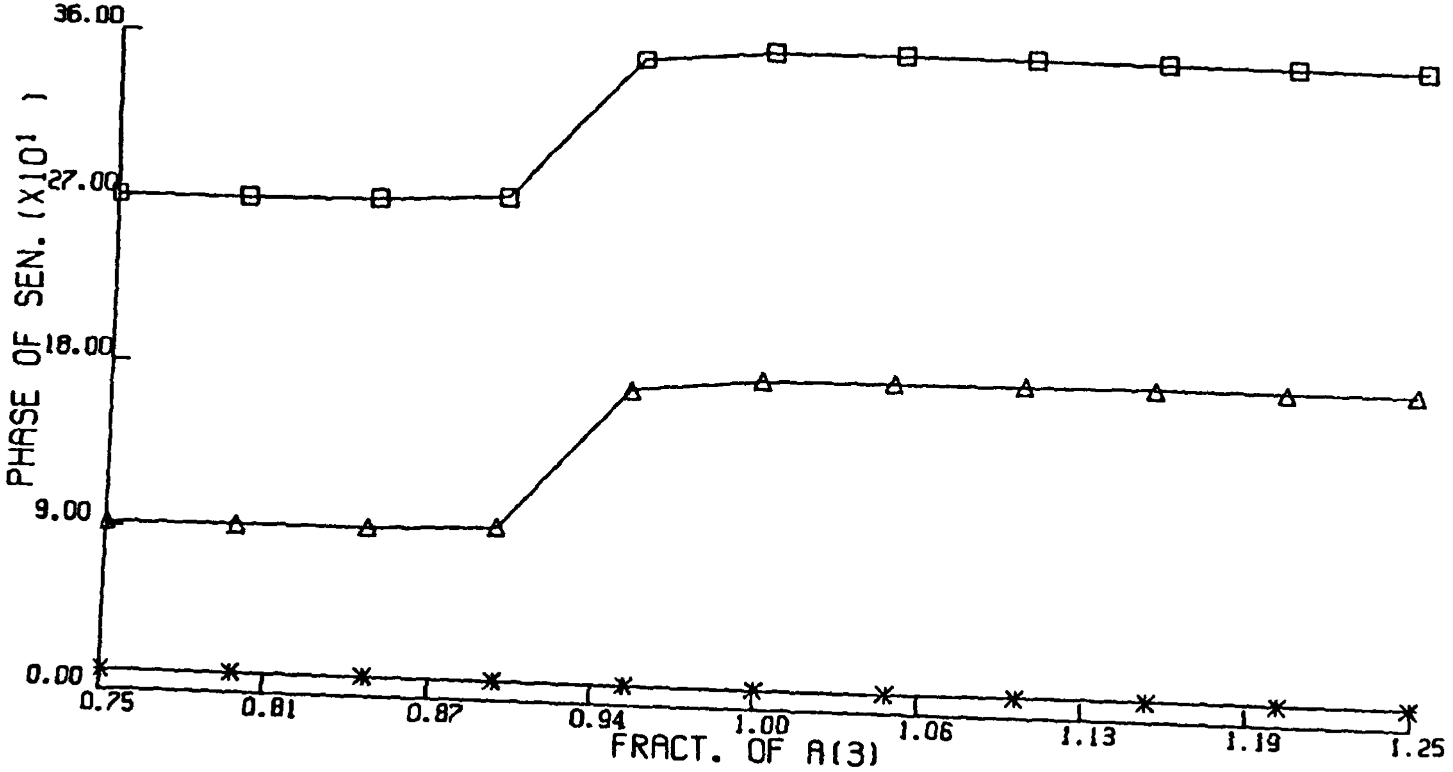


FIGURE 6 - PHONEME IY (1) S(3,*)

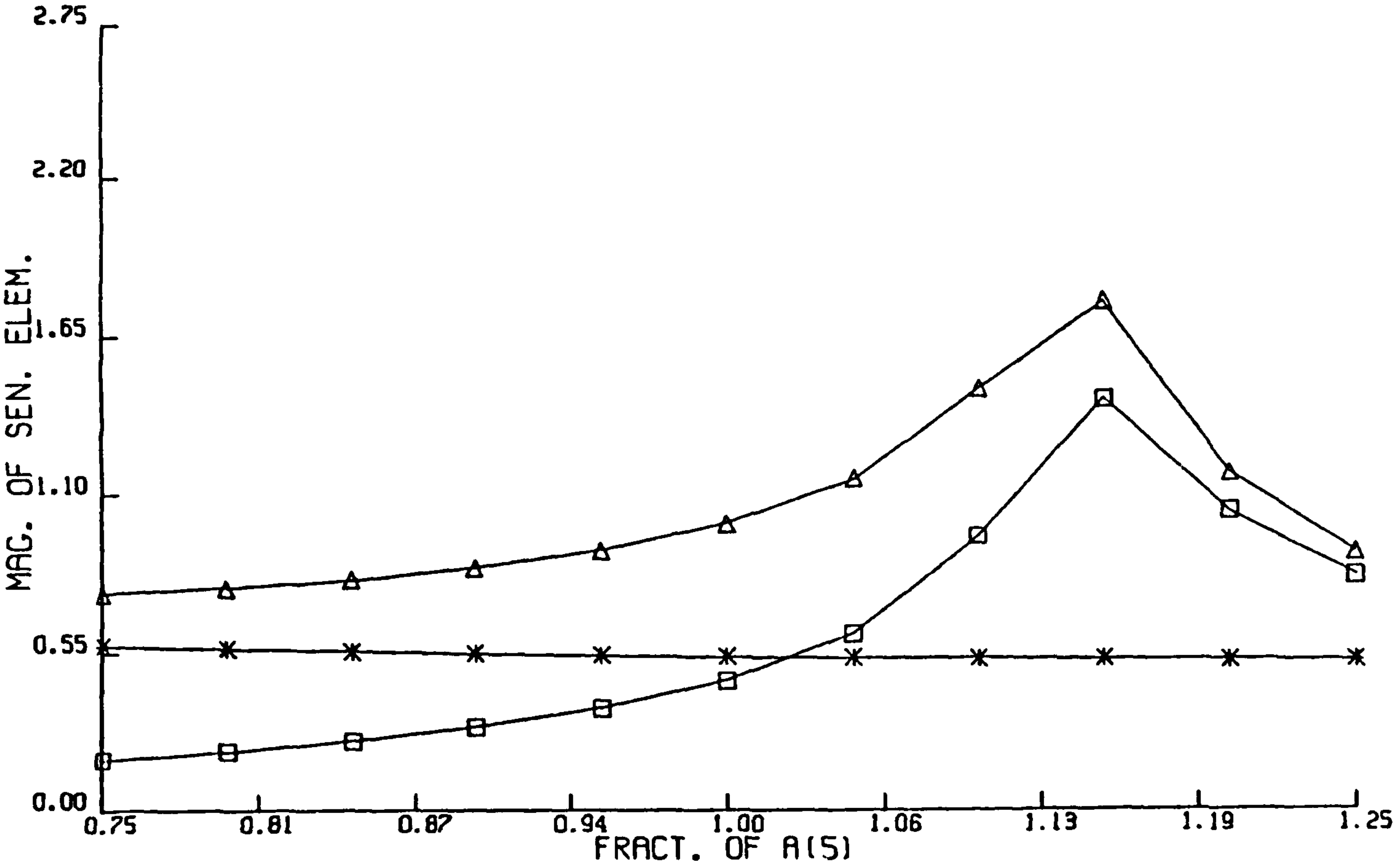


FIGURE 7 - PHONEME IY (1) S(5,x)

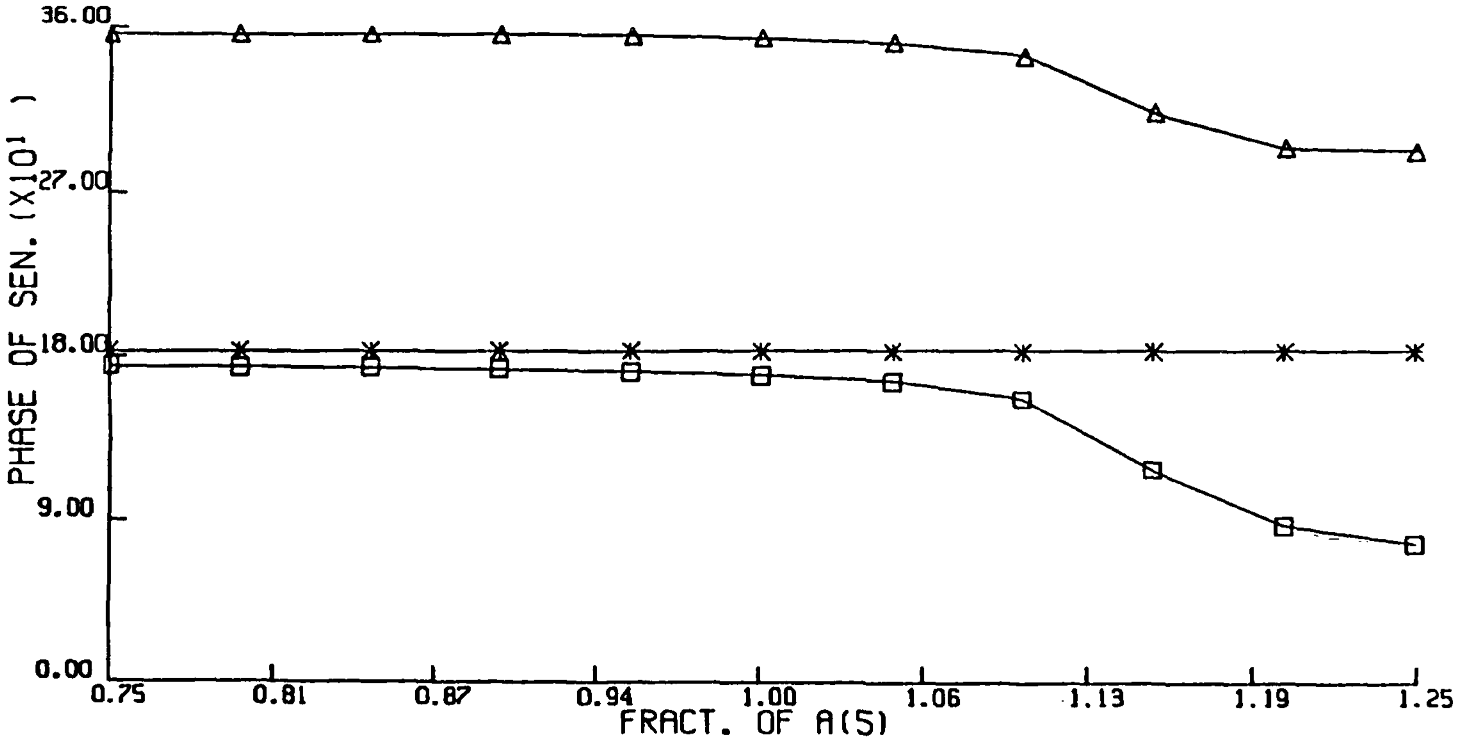


FIGURE 8 - PHONEME IY (1) S(5,x)

The observation that the phase relations described for odd and even numbered coefficients holds for all vowels suggests a categorical indicator for non-vowel waveforms.

Further examination of the phase curves of Figures 6 and 8 show that coefficient a(3) is essentially as low, and coefficient a(5) is essentially as high, as possible without formants 2 and 3 moving "around the corners" in their associated root-locus branches. That is, with nominal coefficient values, formants 2 and 3 are essentially as close together as possible without major changes in their bandwidths. This type of relationship between formants 2 and 3 was found to hold for the mid-vowel /ER/ and all the front vowels /IY, IH, EY, EH, AE/. Thus there is a clear categorical indicator for this group of vowels.

In contrast, the nominal value of coefficient a(3) is intermediate between the root-locus corners of formants 2 and 3 and the root-locus corners of formants 2 and 1 for the low back vowel /AH/. Also, Figures 9 and 10 show that S(5,*) is now quite different in that the nominal value of coefficient a(5) is essentially as low as possible without formants 1 and 2 moving around their root-locus corners. Again, these kinds of relationships hold for all the back vowels /AA, AO, AH, OW, UH, UW, AY, AW, OY/ and serve as a clear categorical indicator.

Changes in the sensitivity elements also reflect the changes that occur in moving from a high front vowel like /IY/ to a low front vowel like /AE/. Sensitivity elements S(3,1) and S(5,1) are larger for vowel /AE/ since the root-locus corners for formants 2 and 3 are closer to the root-locus corners for formants 1 and 2. The changes in sensitivity element S(3,1) in moving from the highest to the lowest front

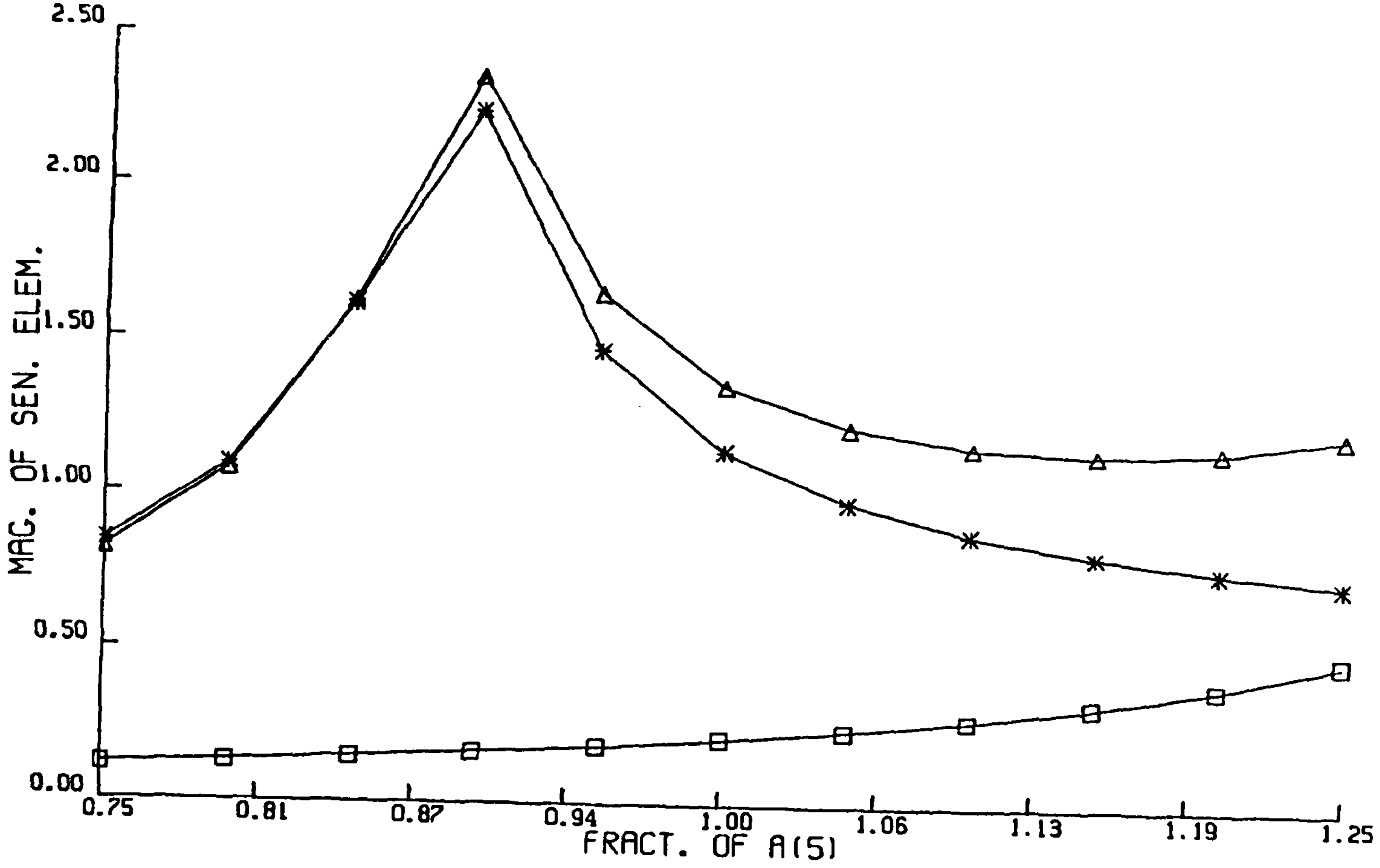


FIGURE 9 - PHONEME AA (6) S(5,*)

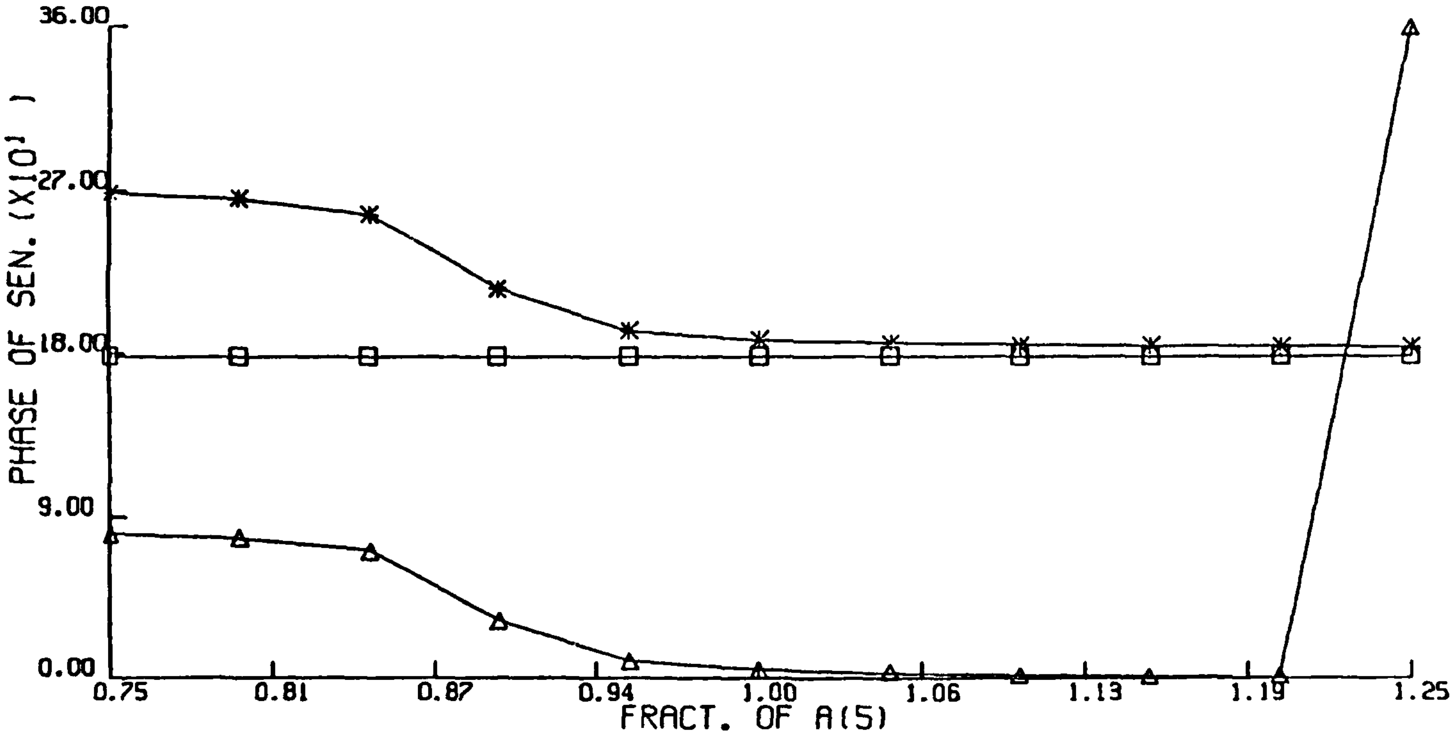


FIGURE 10 - PHONEME AA (6) S(5,*)

vowel are shown in Figure 11. To mimic the effect of noise, also shown on the figure is how the sensitivity element S(3,1) changes due to a ±5% change in coefficient a(5).

These results suggest that the sensitivity elements may be sufficient to identify the particular front vowel. If the starting conditions of some front vowels are similar, such as the long/short pair /EY, EH/, further specificity may be obtained by observing the subsequent changes in the sensitivity elements. For example, the Klatt data is somewhat diphthongized and S(3,1) for phoneme /EY/ decreases from .067 to .023, whereas for phoneme /EH/ it increases from .087 to .150.

The changes that occur among the group of back vowels are reflected by the relative location of the root-locus corners for formants 2 and 3 and the root-locus corners for formants 1 and 2. Values of the sensitivity matrix are a measure of these locations and may be sufficient to identify sub-groups as well as the particular back vowels. Illustrated by Figure 12 is the low back vowel sub-group /AA, AY, AW, AH/. Subsequent changes in the diphthongs /AY, AW/, (△,□) may again provide a greater specificity among the elements of this sub-group. The remaining back vowels are shown in Figure 13 where /OY/ (△) is another diphthong.

For the high front vowel /IY/, Stevens' acoustic analysis predicted a low first formant and that formants 2 and 3 should be close together. The sensitivity analysis of Klatt's data corroborates and extends Stevens' results by showing that this condition holds for all front vowels and for the mid-vowel /ER/. Stevens also considered the low and high back vowels /AA, UW/ and in each case concluded that formants 1 and 2 should be close. Again, the sensitivity analysis of Klatt's

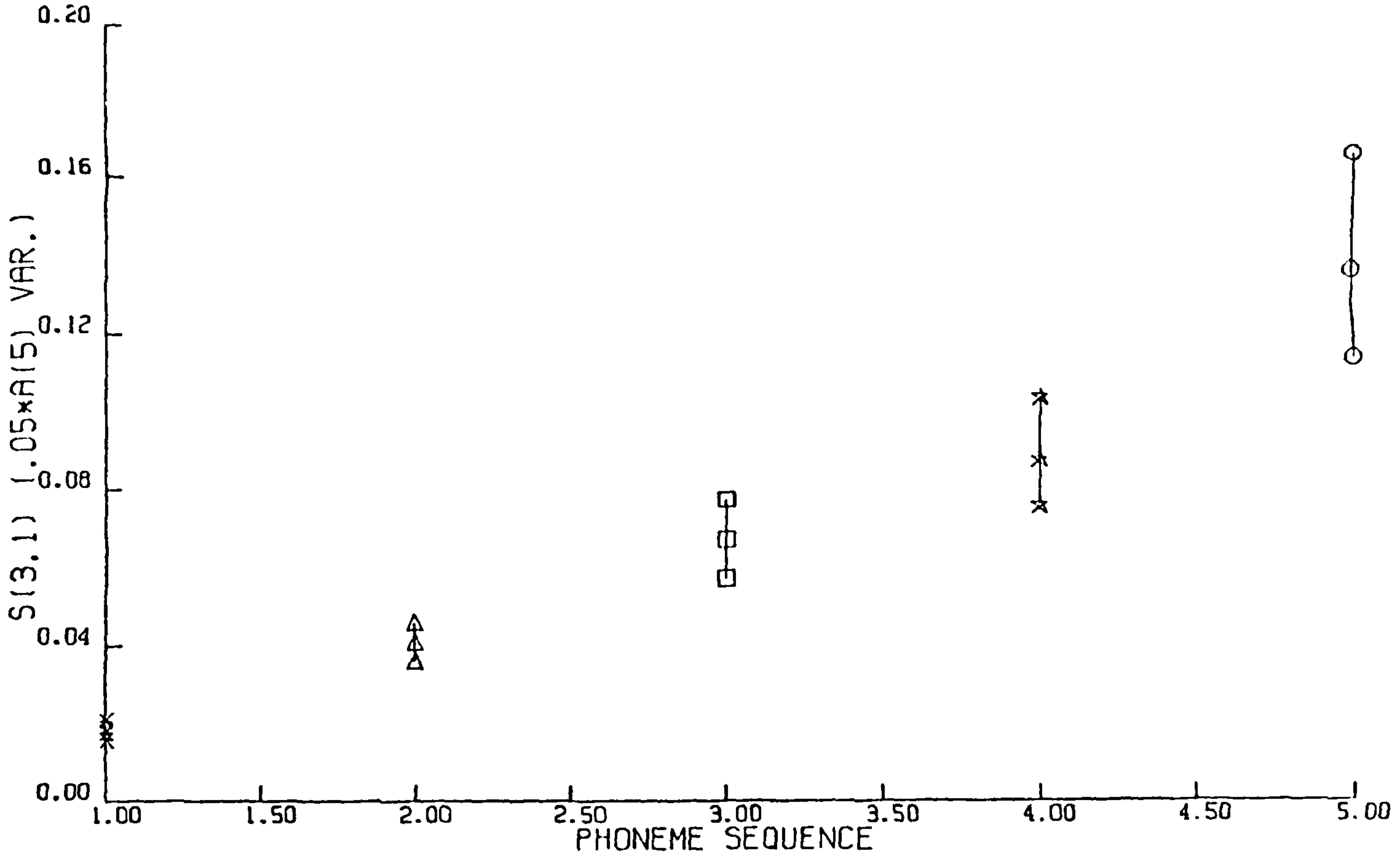


FIGURE 11 - KLATT DATA /IY.IH.EY.EH.AE/

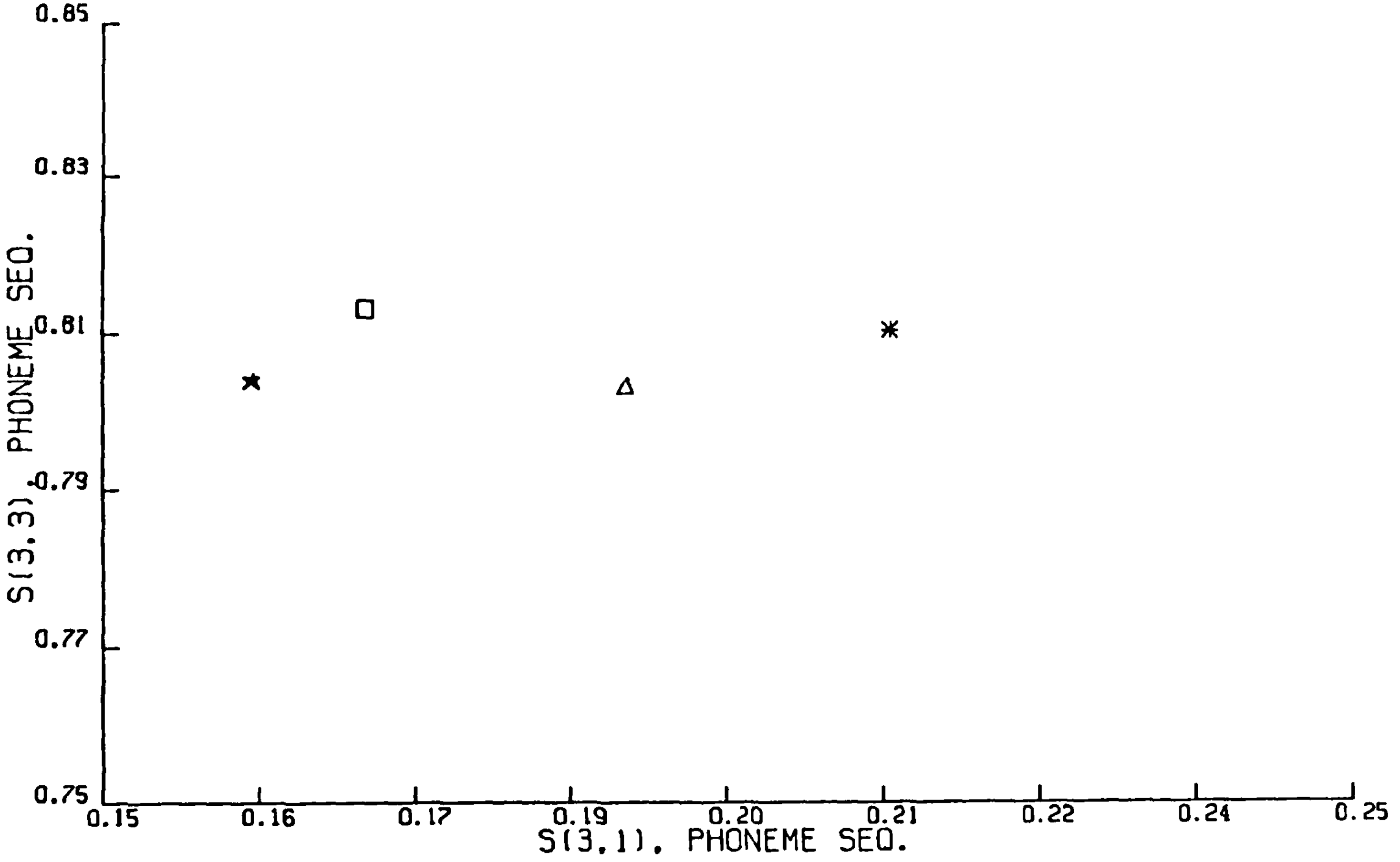


FIGURE 12 - KLATT DATA /AA,AY,AW,AH/

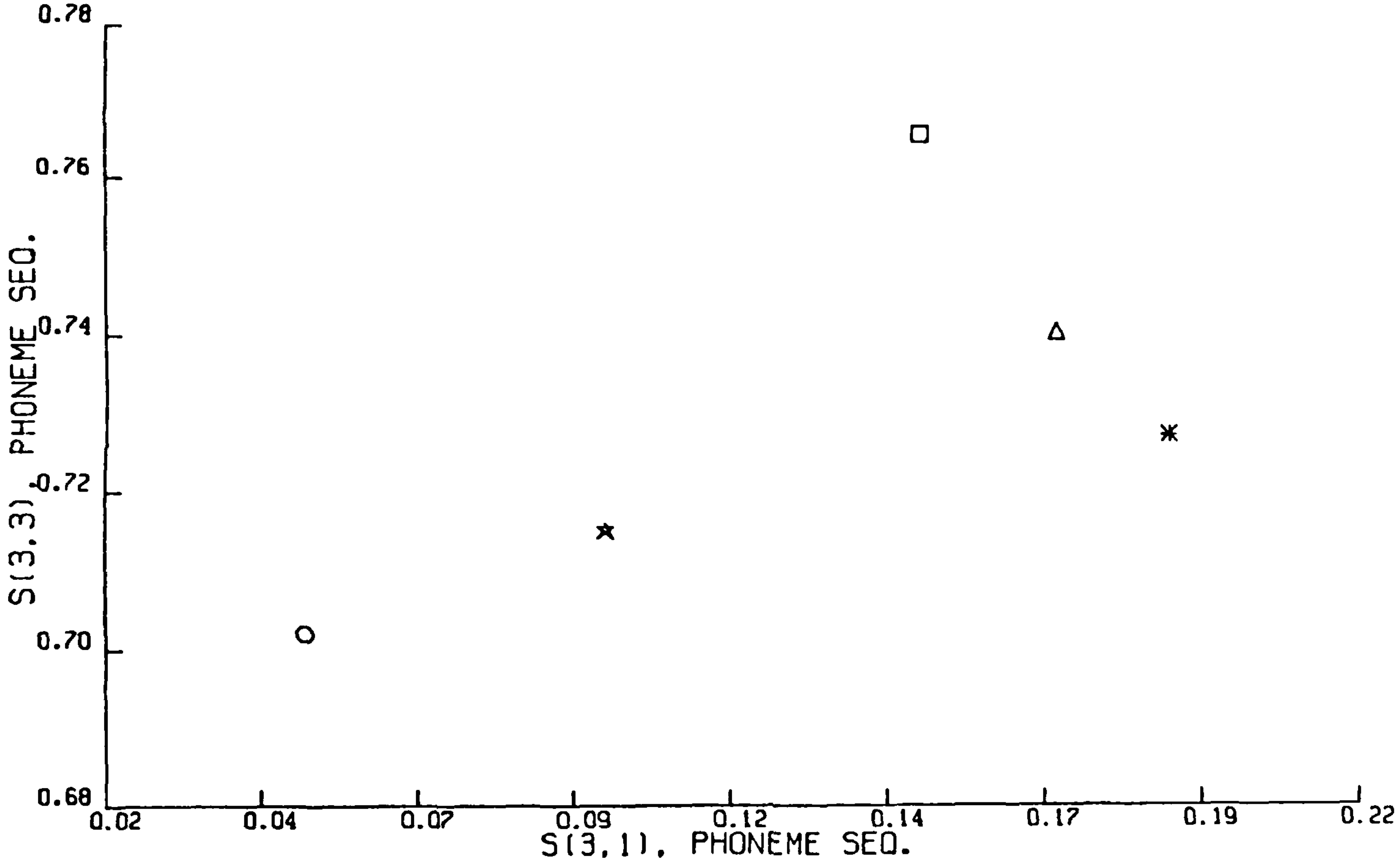


FIGURE 13 - KLATT DATA /AO,OY,OW,UH,UW/

data corroborates and extends Stevens' results by showing that this condition holds for all back vowels.

In summary, the sensitivity matrix was evaluated for the initial and final representations of the fifteen vowels in Klatt's data set. It was found that (1) the sensitivity matrix does provide a measure of the degree to which a sound is "on target" by locating the sound relative to the root-locus corners of formants 2 and 3 and those for formants 2 and 1; (2) that formants 2 and 3 being close to their root-locus corners provides a categorical indicator for the group of front vowels; (3) that formants 2 and 1 being close to their root-locus corners provides a categorical indicator for the group of back vowels; and (4) that particular elements of the sensitivity matrix may provide sufficient information to identify the particular vowel.

V. PRESENT WORK:

Overview.

The positive results obtained with the Klatt data demonstrated that further investigations with the sensitivity analysis were warranted. A sequence of studies utililzing "real" speech obtained from several male and female speakers should clarify its usefulness. Furthermore, any particular method of segmentation and identification of phonemes should be challenged by speech material which presents, in both a controlled and naturalistic manner, as many of the factors known to cause acoustic-phonetic variations as possible. The entire set of English vowels should be used in conjunction with a number of consonants that sample coarticulatory variations. These consonants should include (1) differing manners - stops, fricatives, approximants, liquid vs. glide vs. nasal contrast, (2) different voicing, and (3) differing place-labial vs. velar. Stress, tempo, and word position-structure should also be included.

The sequence of studies could be described as follows. Task I should evaluate the changes in the sensitivity matrix that occur during the production of the vowels in the words, test the conclusions reached with Klatt's data set, and evaluate inter-speaker variations. Task II should evaluate and interpret changes in the sensitivity matrix for the vowels due to coarticulation. Task III should evaluate and interpret the sensitivity matrix for the differing initial consonants. Finally, Task IV should use the above results to build a reference library, and should evaluate the efficacy of the sensitivity matrix in terms of the accuracy of the resulting phonetic representation of unknown speech.

It was considered premature and unrealistic to include all these factors and studies in the current research plan. Instead, only the three stop-consonants /b, d, g/ were used in single words with the six vowels /i, ɛ, æ , ɔ , ʌ , u/. The three consonants were selected because they have the same manner, the same voicing, but the differing place should induce substantial coarticulatory variations. Of the six vowels, three are front and three are back. They were selected because /i/ is "far away" from /ɛ, æ /, whereas /ɛ/ and /æ/ are "close" and "difficult" to distinguish using current methods and techniques. The same kind of relationship holds for /u/ and the pair /ɔ , ʌ /. The words shown in Table 1 were chosen to express these sets of consonants and vowels.

TABLE 1. TEST WORDS USED IN CARRIER PHRASE "SAY (WORD) AGAIN."

VOWEL	CONSONANT		
	/b/	/d/	/g/
/i/	bead	deed	geese
/ε/	bed	dead	guess
/æ/	bad	dad	gas
/ʌ/	bud	dud	gus
/ɔ/	baud	dawdle	gauze
/u/	booed	dude	goose

Using this selected set of data, the studies were limited to those of Task I. It was anticipated that the elements of the sensitivity matrix would change during the production of a test word. These changes in the sensitivity matrix were analyzed, as with Klatt's data set, to determine if (1) there were general properties that hold for all vowels and thereby provide a measure of the degree to which a vowel was "on target," (2) it had properties that provided categorical indicators for particular subgroups of the vowels, and (3) whether they provided sufficient information to identify each vowel. Also, multiple repetitions and multiple speakers allowed modest statistical assessment of intra and inter-speaker variations.

Methods.

The subjects were five male speakers with differing fundamental frequency and dialect. Dialects chosen were representative of Vermont, Pennsylvania, Virginia, Rhode Island, and Michigan. Speakers with these dialects were readily available and sampled those described by the Linguaphone Institute, American Dialect Series. Each speaker made, in fully randomized order, three repetitions of each word in the carrier phrase "Say (word) again" on each of four days. Thus the final corpus of utterances consisted of 5 speakers X 12 repetitions X 6 vowels X 3 consonants.

All master recordings were made in an Industrial Acoustics sound room using a Nakamicki 550 portable cassette recorder. Subjects were instructed to use equal effort to produce the samples while speaking in a normal conversational tempo and voice into a head-band held Teledyne EC-101 electret microphone. Subjects were instructed to produce a word which was fully pronounced; that is, no casual speech alternations of word structure were accepted. VU levels were monitored as a check on speaking level. Two expert phoneticians monitored speech productions and rejected any sample which was not jointly recognized "live" as an adequate production of the word. They transcribed each vowel production to determine its perceptual quality vis a vis a traditional phonetic vowel quadrangle.

During playback, the master recordings were bandlimited to 4.8 KHZ and were digitized at a sampling interval of 83 microseconds using the 12-bit A/D converter on the PDP 11-34 computer. Using a waveform editing program, a particular carrier phrase was displayed on the AED 512 color graphics terminal, and the test word was excised for storage and analysis. Each test word was divided into successive frames. For each frame, the formant frequencies and bandwidths were calculated via linear predictive analysis.[19-23] Using only the first three formant frequencies and/or bandwidths for each frame, the elements of the sensitivity matrix were calculated. As an aid for interpretation of results, the color graphics terminal was used to plot the test word waveform along with waveforms of the corresponding elements of the sensitivity matrix. The following discussion develops the signal analysis problems and methods in greater detail.

In linear predictive analysis, an all-pole model of a signal is determined by predicting each signal sample as a linear combination of some number of previous signal samples. Programs were implemented on the laboratory computer, a PDP 11-34, for both the covariance and autocorrelation methods of determining the linear predictive coefficients.[24] After testing each method with typical sets of speech data, it was decided that further studies should utilize the covariance method since it provided less frame-to-frame variation and a smaller prediction error than did the autocorrelation method.[25]

These initial studies with typical speech data were also used to decide on other details of the signal analysis methods. By comparing the results obtained when the predictor order was varied from 8 to 22, it was found that the predictor order should be at least 16. The resolution provided by this high predictor order was necessary when, as in high back vowels like /UW/, two formant frequencies of unequal bandwidth were close together. Use of high-pass filtering, where the corner frequency is below the first formant frequency, and/or use of pre-emphasis would allow a smaller predictor order.[26] Because of their added complexity and their influence on the location of the spectral peak for the first formant frequency, it was decided not to use high-pass filtering or pre-emphasis.

Contiguous and fixed frame lengths of 256 speech signal samples were utilized in the initial studies. Since the frames were not pitch synchronous, the corresponding sensitivity analysis showed some cyclic frame-to-frame variations. Thus, it was decided to adopt a quasisynchronous framing method where the frame length was selected as twice the estimated pitch period and each frame began in the middle of a pitch period. Pitch period estimates were made using a simplified filter tracking algorithm (SIFT) by Markel[27] and were updated each 512 speech samples. In this algorithm, the speech waveform was down-sampled, low-pass filtered, and represented by a fourth order all-pole model. The inverse filter was formed by inverting the transfer function of the all-pole model. The pitch period was estimated from the peak in the autocorrelation sequence which was calculated with the signal that results from passing the processed signal through the corresponding inverse filter.

After framing the speech data, as described above, and calculating the coefficients of the all-pole model using the covariance method, it was necessary to calculate the formant frequencies and their bandwidths. Initial studies utilized the roots of the denominator of the transfer function.[20] Any real roots, or any complex-conjugate roots with "wide" bandwidths were discarded since the resonant peaks were considered to be represented by complex-conjugate poles with "narrow" bandwidths. This method was abandoned because of difficulty in judging narrow versus wide bandwidths. Instead, the all-pole model was used to calculate the signal spectrum, and a peak-picking algorithm was implemented.[26] It was necessary to develop some decision logic since multiple spectral peaks were sometimes found to occur near or below the first formant frequency. This was done by placing upper and lower bounds on the first formant frequency and associating it with the spectral peak of the smallest bandwidth that occurred between these bounds.

Once the first three formant frequencies and bandwidths were determined for each frame, the elements of the sensitivity matrix were calculated using equations (5) and (3). It was anticipated that the elements of the sensitivity matrix would change during the production of a test word and that the angle of the sensitivity elements would provide a measure of the degree to which a vowel was "on target" by characterizing the root locations relative to the root-locus corners. Results from initial studies tended to support this hypothesis but they also showed frame-to-frame variations in sensitivity angle that resulted from corresponding changes in the formant bandwidth estimates. Regardless of whether these variations occurred because the signal analysis methods did not provide accurant formant bandwidth estimates and/or because the speech process does not accurately control energy loss mechanisms, it was decided to approximate the system by a lossless model.

The hypothesis of a lossless model implies that knowledge of the formant frequency bandwidths is not essential for the recognition of the vowels in the test set of words. Under these conditions, equations (5) and (6) become:

$$r_i = j2\pi f_i \qquad i = 1, \ldots, \frac{n}{2} \tag{7}$$

$$q(s) = s^n + a_3 s^{n-2} + \ldots + a_{n-1}s^2 + a_{n+1} \tag{8}$$

where n is even.

In order for the angle of the sensitivity elements to be expressed in rectangular cartesian form, the definition of equation (2) was changed as follows:

$$S(k,i) = \frac{a_k}{|r_i|} \frac{dr_i}{da_k} \tag{9}$$

This definition easily leads to the following closed-form expression:

$$S(k,i) = - \frac{a_k}{|r_i|} \frac{\left(-r_i\right)^{n-k+1}}{\left.\frac{dq(s)}{ds}\right|_{s = -r_i}} \tag{10}$$

Equation (10) was used in lieu of equation (3) because it was computationally more efficient.

Summing the sensitivity elements across any row of the matrix easily leads to:

$$\sum_{i=1}^{n} S(k,i) = \begin{cases} 0 & \text{if } 1 \neq n+1 \\ 1 & \text{if } k = n+1 \end{cases}$$

Combining this constraint, on the elements in any row, with the fact that the sensitivity elements of complex-conjugate roots r_i are themselves complex-conjugates, shows that only $\frac{n}{2} - 1$ of the elements in each row are independent. Furthermore, equation (10) shows that the elements in each column are simply related by the factor $a_k \ (-r_i)^{-k}$. Thus, $\frac{n}{2} - 1$ of the elements in any row of the sensitivity matrix should be sufficient to characterize the root sensitivity patterns. For this research project, n was equal to 6 since only the first three formants were considered. Thus, attention was focused on the two elements S(3,1) and S(3,3) which expressed the sensitivity of formant frequencies 1 and 3 to changes in coefficient a_3.

For this lossless case, the corresponding root-locus of Figure 1 is simplified as shown in Figure 14. Since only the portion of the root-locus on the $j\omega$ axis is consistent with the lossless condition, the points marked CR2 and CR1 portray the minimum, $a_{3\ min}$, and maximum, $a_{3\ max}$, values permitted for coefficient a_3. At corner location CR2, formants 2 and 3 are equal and attain their corresponding maximum and minimum values. The minimum value for the frequency of formant 1 occurs with $a_{3\ min}$. At the other corner location CR1, formants 1 and 2 are equal and attain their corresponding maximum and minimum values. The maximum value for the frequency of formant 3 occurs with $a_{3\ max}$. A direct calculation was made for these values of coefficient a_3 and the corresponding frequencies.

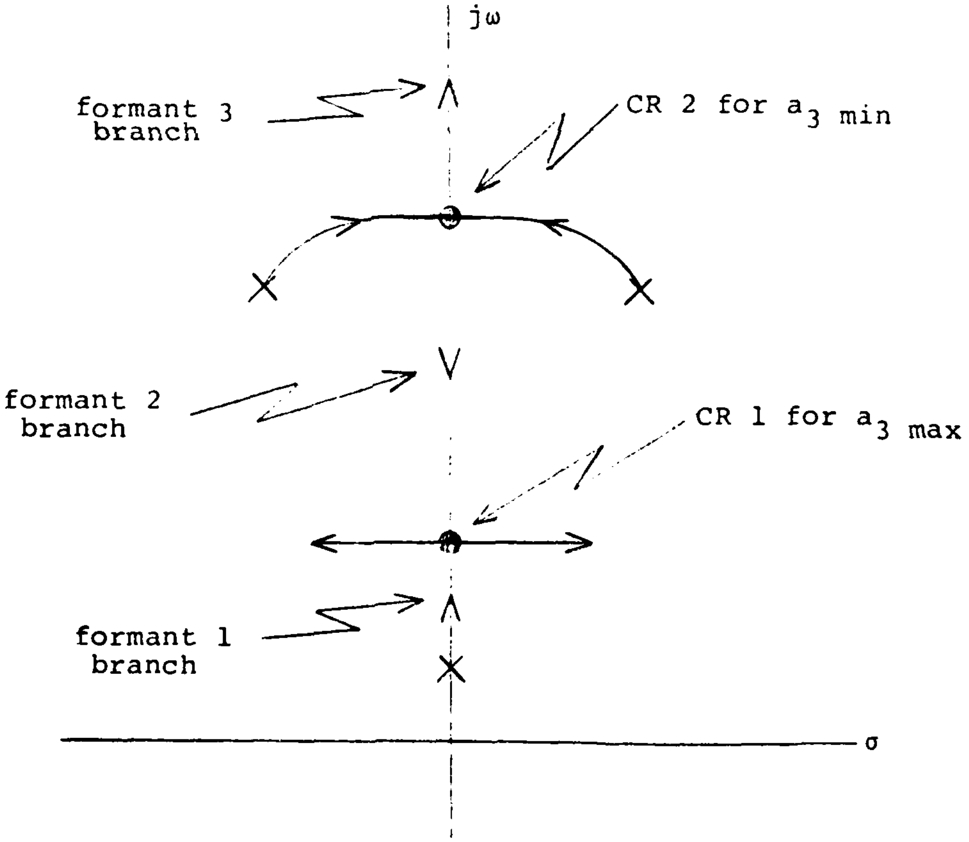


FIGURE 14 - ROOT-LOCUS SKETCH FOR LOSSLESS MODEL

It was anticipated that the degree to which a vowel was "on target" could be described by the root locations relative to the root-locus corners CR2 and CR1. Two measures of this property were defined:

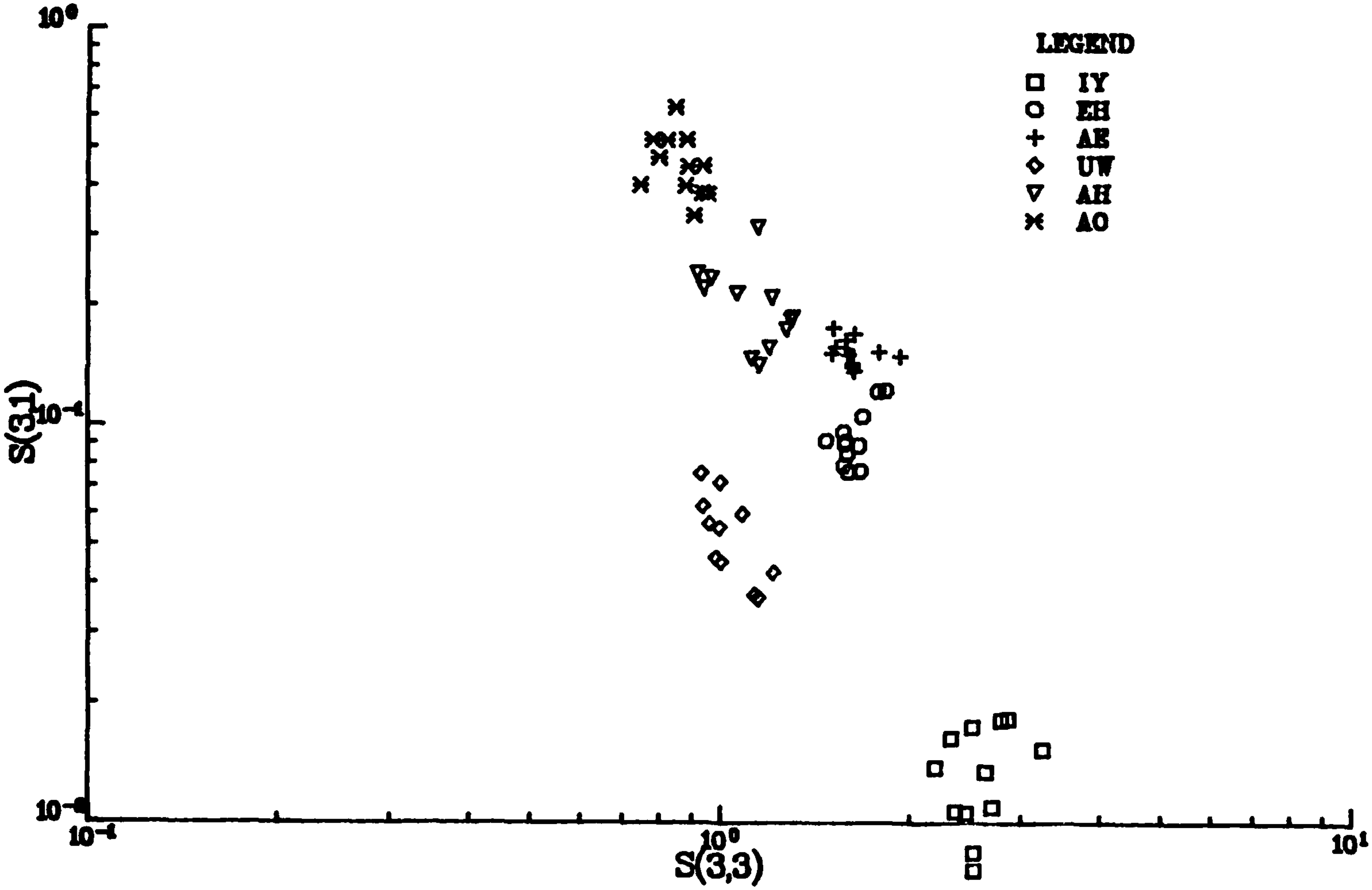
$$\text{Corner Ratio 1} = \frac{a_3}{a_{3\ \max}} \tag{11}$$

$$\text{Corner Ratio 2} = \frac{a_{3\ \min}}{a_3} \tag{12}$$

If formants 2 and 3 being close was characteristic of front vowels, then Corner Ratio 2 should be large. Likewise, if formants 1 and 2 being close was characteristic of back vowels, then Corner Ratio 1 should be large.

Initial studies indicated that from frame to frame, Corner Ratio 1 was a smooth and well behaved curve during the vowel portion of a word and could thus serve to segment the vowel interval. Moreover, within the vowel interval, the curve generally had "flat" spots. A computer program was written, called Algorithm 1, to locate the maximum such "flat" spot so that its utility as a frame selector could be evaluated. For some speakers, as illustrated by Figure 15, the sensitivity elements S(3,1) and S(3,3) of the selected frame may be used to accurately identify each of the six vowels.

Each vowel /IY, EH, AE, AH, AO, UW/ is indicated by a different symbol, as shown in the legends, and there is no overlap among these six vowel groups. For other speakers, the sensitivity values of the frames selected by Algorithm 1 may overlap for neighboring vowels such as /EH, AE/. In such cases, the corresponding vowel may only be identified as one of a neighboring pair of vowels. This result suggested a two



SUBJECT #5

FRAME SELECTED VIA ALGORITHM 1

FIGURE 15 .

level identification scheme, where the second level was based on vowel-pair-specific algorithms that made use of both Corner Ratios. Two such algorithms, called Algorithm 2, were developed: to distinguish /EH/ from /AE/ and to distinguish /AO/ from /AH/. These algorithms are not speaker specific, nor are any of the other algorithms.

For the neighboring front vowels /EH, AE/, Algorithm 2 selects the frame at the maximum "flat" or "smooth" region of the Corner Ratio 2 curve that occurs prior to the maximum value in the Corner Ratio 1 curve. Thus, at the selected frame, formants 2 and 3 are close. In the case of neighboring back vowels /AH, AO/, Algorithm 2 selects the frame (1) at a minimum of the Corner Ratio 2 curve (2) at a maximum "flat" spot if the Corner Ratio 2 curve trends upward or (3) at a minimum "flat" spot if the Corner Ratio 2 curve trends downward. Thus, at the selected frame, formants 2 and 3 either have a maximum spread or have a spread and stationary relationship.

Algorithm 2 has been applied "by hand" to obtain the results presented in this report. For use in future studies, a computer program is being written that will implement Algorithm 2. Other developments needed in the area of signal processing are outlined in the section titled Future Work.

Results.

The five speakers used in this study were chosen to represent a variety of American English dialects as defined by Thomas.[28] Speaker 1, from rural Vermont, represented the Eastern New England area. Speaker 2, originally from Detroit, represented the North Central region. Speaker 3, from Rhode Island, represented the New York area. Speaker 4, from coastal Virginia, represented the Mid Atlantic area. Speaker 5, from Pittsburgh, represented the Western Pennsylvania area. Table 2 shows the transcription most commonly used to describe each speaker's production of general dialect variants of each vowel as judged "live" by two expert phoneticians.

TABLE 2. GENERAL AMERICAN ENGLISH DIALECT PRODUCTION AND ASSOCIATED SPEAKER PRODUCTS

	GENERAL AMERICAN DIALECT (G.A.D.)					
SPEAKER	/i/	/ɛ/	/æ/	/ʌ/	/u/	/ɔ/
1	i	ɛ	æ	ʌ	u	a*
2	i	ɛ	æ	ʌ	u	ɔ
3	i	ɛ	æ	ʌ	u	ɒ *
4	i	ɛ^{æ} *	æ^{ɛ}*	ʌː*	u	ɔ
5	i	ɛ	æ	ʌ	u	ɔ

*Productions which differ from G.A.D.

Speakers 2 and 5 consistently maintained G.A.D. pronunciations of all vowels. Speaker 1 used a consistent /a/ or /a^{ɔ}/ production for /ɔ/. Speaker 3 used a consistent /ɒ/ production for /ɔ/. Speaker 4 demonstrated a widespread tendency to diphthongize all productions, particularly productions of /ɛ/, /æ/, and /ʌ/.

The results described below were obtained from the analysis of one of the three repetitions of each test word recorded on each of the four days. The data from the remaining repetitions remains available for future studies. As described in the Methods section, each test word was divided into successive frames. For each frame, the coefficients of an 18 pole model were calculated using the covariance method. A peak-picking algorithm located the first 3 formant frequencies from the corresponding signal spectrum. These were used, under the hypothesis of a lossless model, to calculate Corner Ratio 1, Corner Ratio 2, and the sensitivity elements S(3,1) and S(3,3) for each frame.

Using data from all six vowels of Subject 2, Figures 16 through 23 show example plots of these four quantities versus Sequence Number, which is the speech sample number. For ease of comparison, the three front vowels are grouped in each plot as are the three back vowels. In each of these curves, the well behaved regions correspond to the vowel portions of the words. By placing limits on the values of Corner Ratio 1 and Corner Ratio 2 and on their smoothness, the vowel intervals have automatically been segmented as labeled by s, for start, and e, for end.

Figures 16 through 23 also show, as labeled by 1 and 2, the frames selected by Algorithms 1 and 2. As described in the Methods section, and illustrated in Figures 16 and 18, Algorithm 1 located the frame at the maximum "flat" or "smooth" region of the Corner Ratio 1 curves. If the sensitivity values of the selected frame indicate that the vowel can only be identified as one of a neighboring pair of vowels, then a vowel-pair specific form of Algorithm 2 may be utilized. For the EH or AE vowels, Algorithm 2 located the frames at the maximum "flat" or "smooth" region of the Corner Ratio 2 curves, as illustrated in Figure 17. For the contrasting case of the AH or AO vowels, Algorithm 2 located the frame at the minimum of the Corner Ratio 2 curves, as illustrated in Figure 19.

For each test word of each speaker, the sensitivity elements S(3,1) and S(3,3) of the frame selected by Algorithm 2 are plotted in Figures 24 through 28.

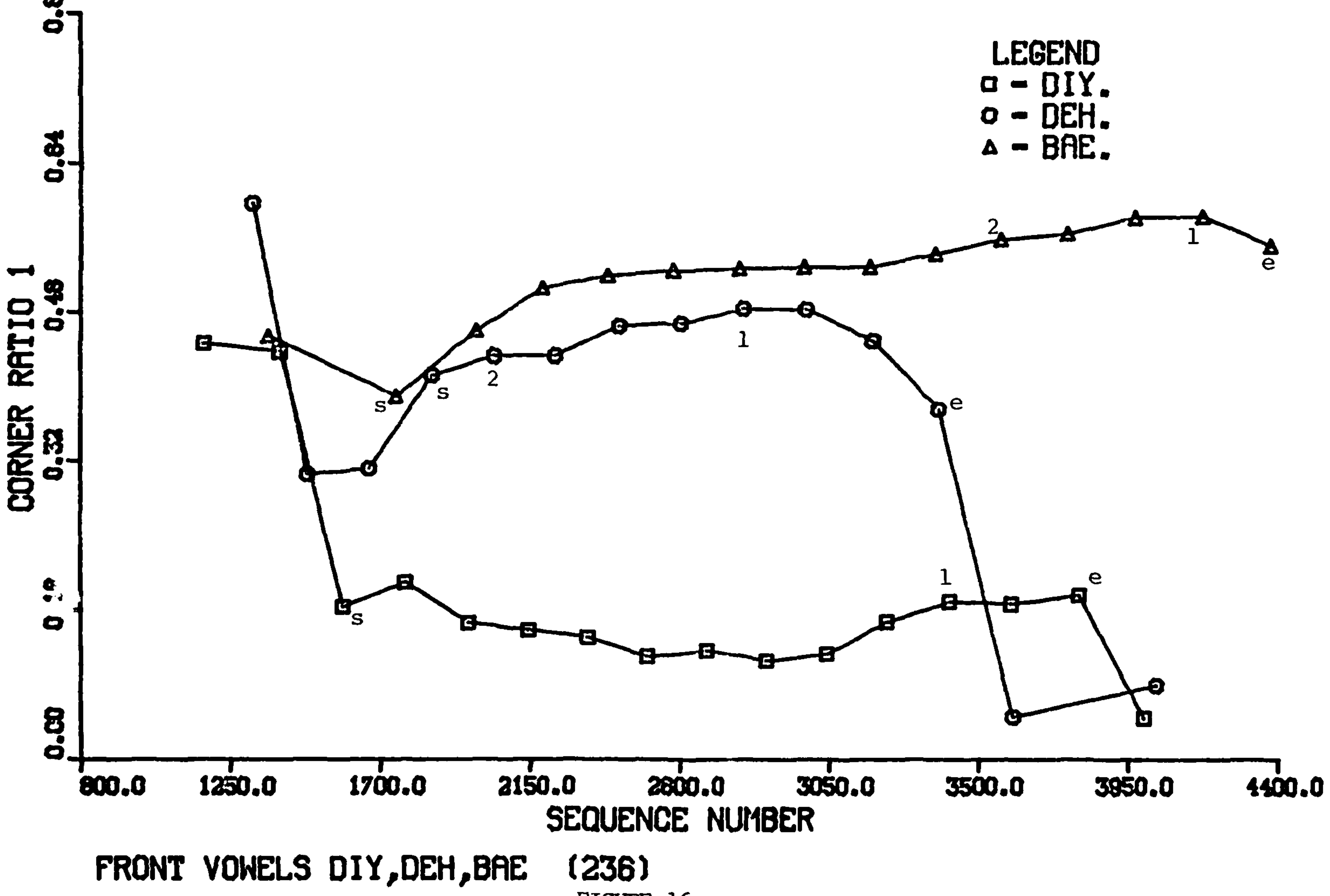

LEGEND
□ - DIY.
○ - DEH.
△ - BAE.
CORNER RATIO 1
0.00
0.16
0.32
0.48
0.64
0.80
800.0
1250.0
1700.0
2150.0
2600.0
3050.0
3500.0
3950.0
4400.0
SEQUENCE NUMBER
s
s
s
1
1
1
2
2
e
e
e
FRONT VOWELS DIY,DEH,BAE (236)

FIGURE 16.

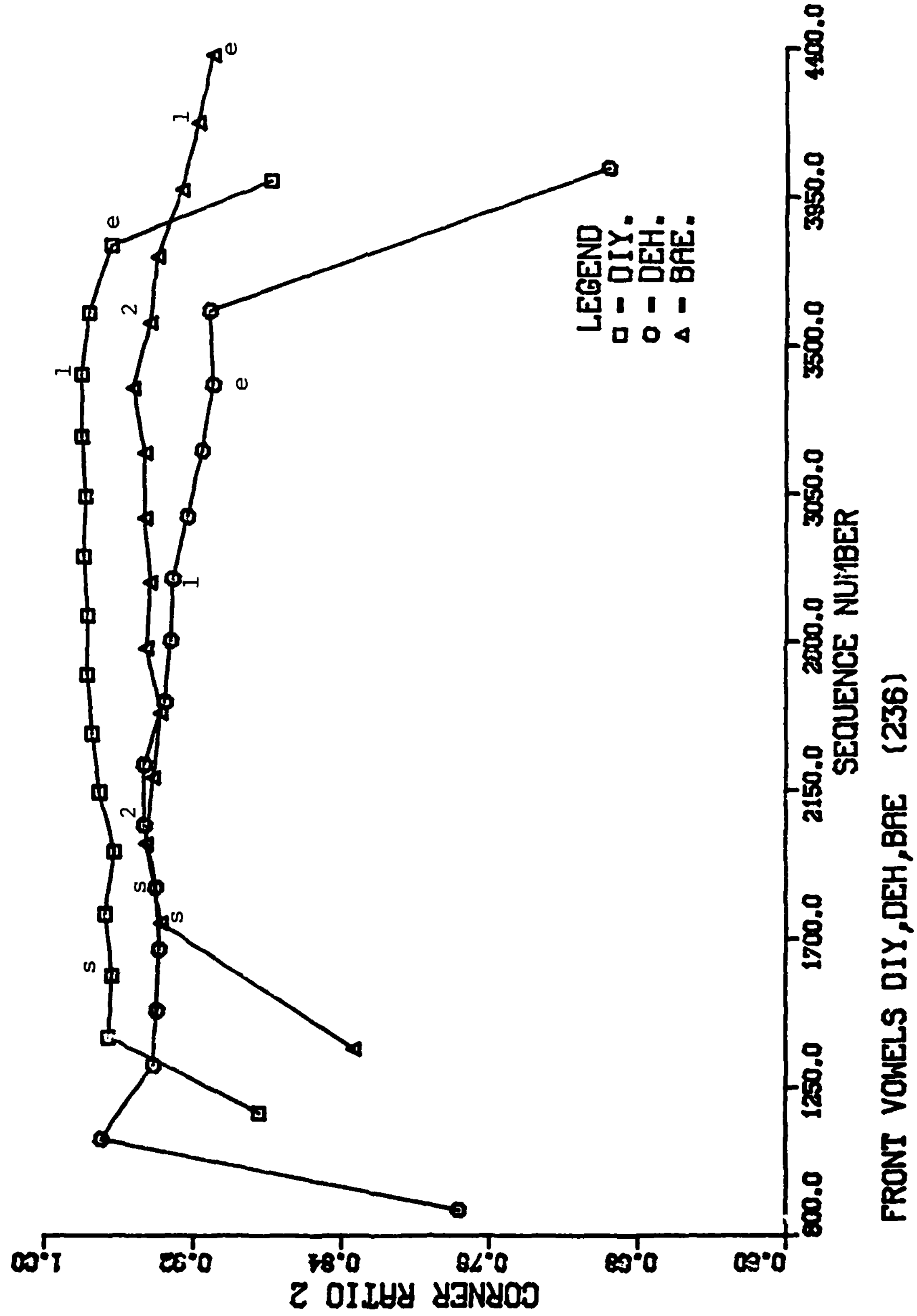
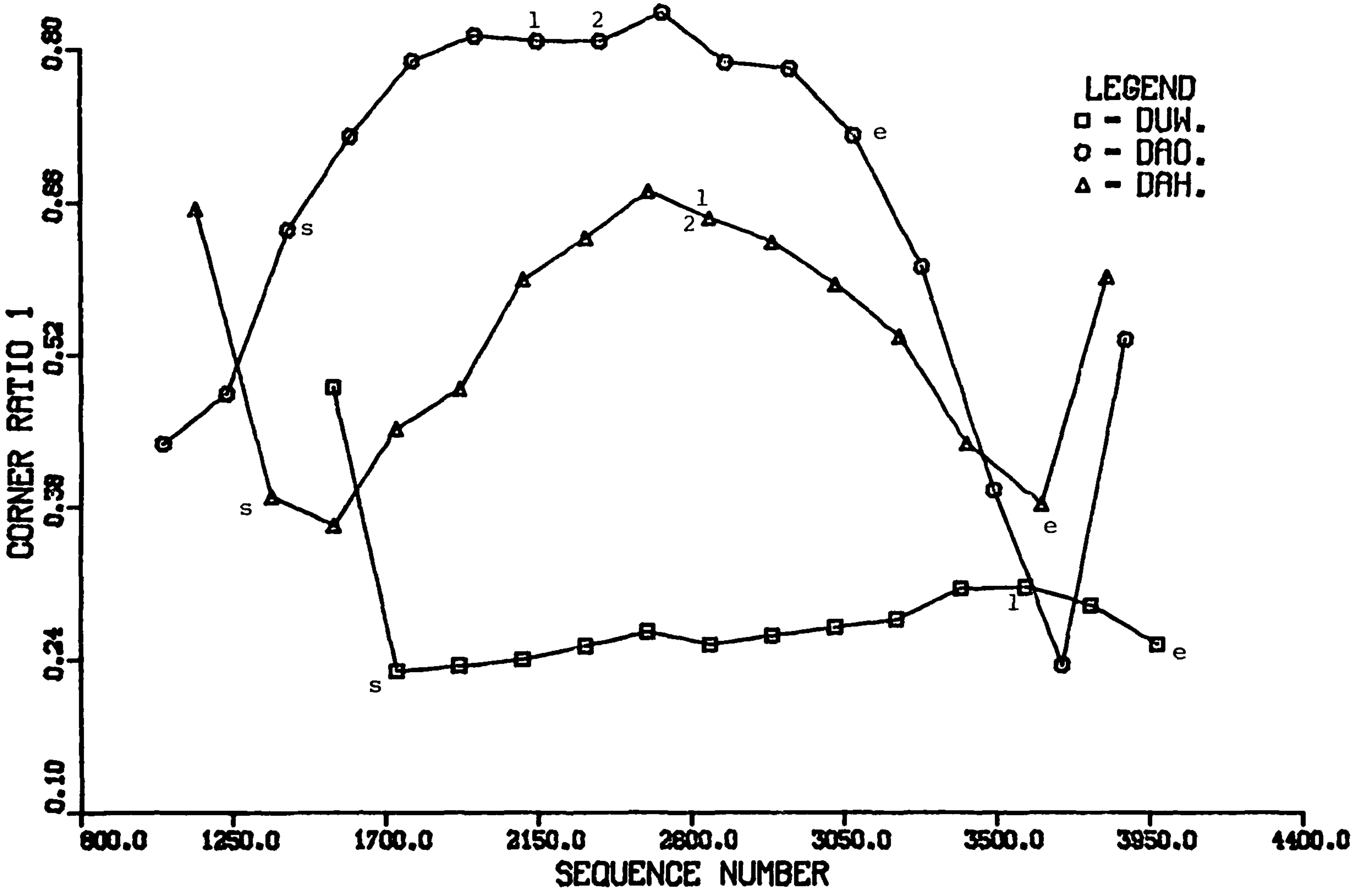

FRONT VOWELS DIY,DEH,BAE (236)
SEQUENCE NUMBER
CORNER RATIO 2
800.0
1250.0
1700.0
2150.0
2600.0
3050.0
3500.0
3950.0
4400.0
0.60
0.68
0.76
0.84
0.92
1.00
LEGEND
□ - DIY.
○ - DEH.
△ - BAE.
s
s
s
2
2
1
1
1
1
e
e
e

FIGURE 17.



BACK VOWELS DUW,DAO,DAH (236)

FIGURE 18.

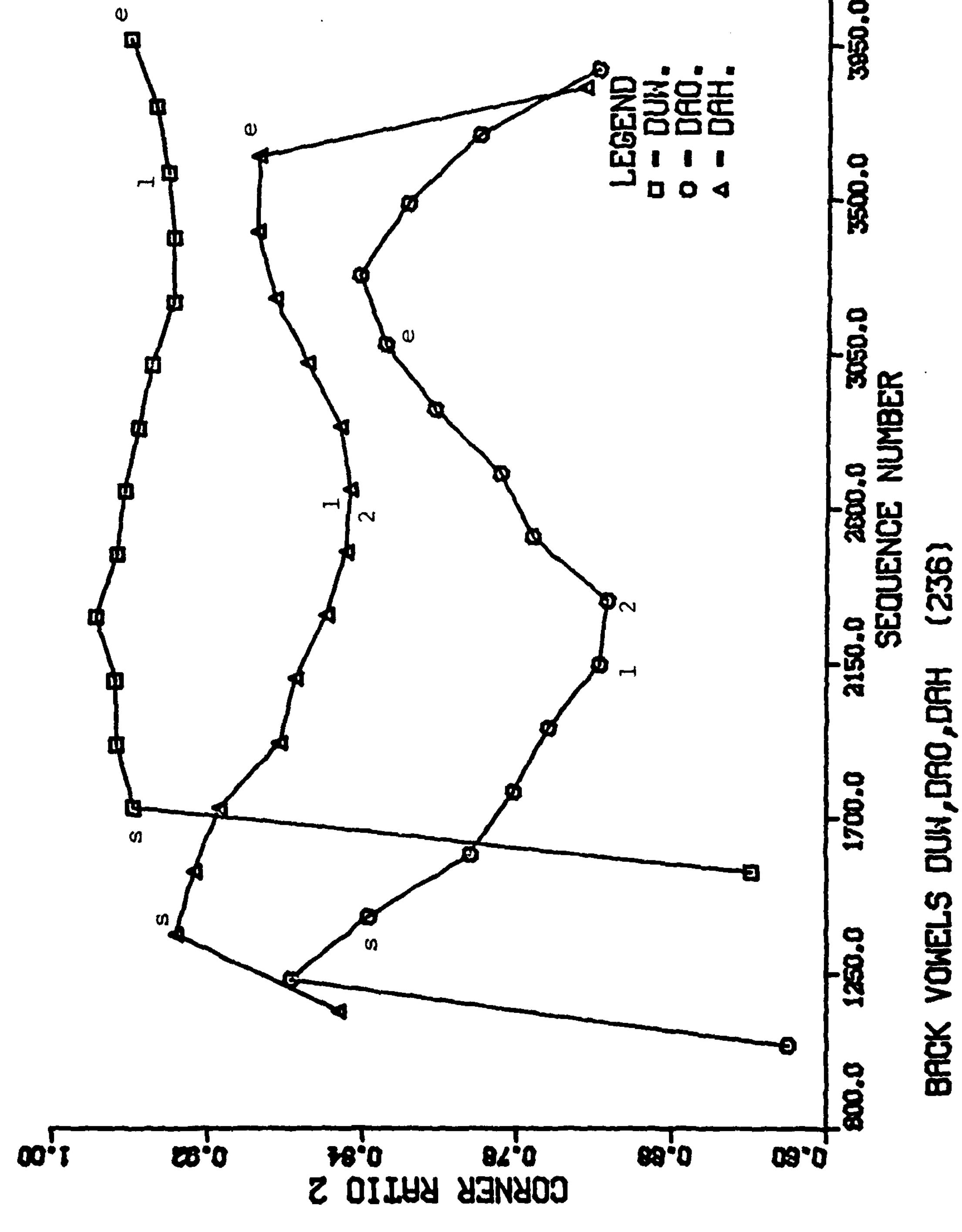
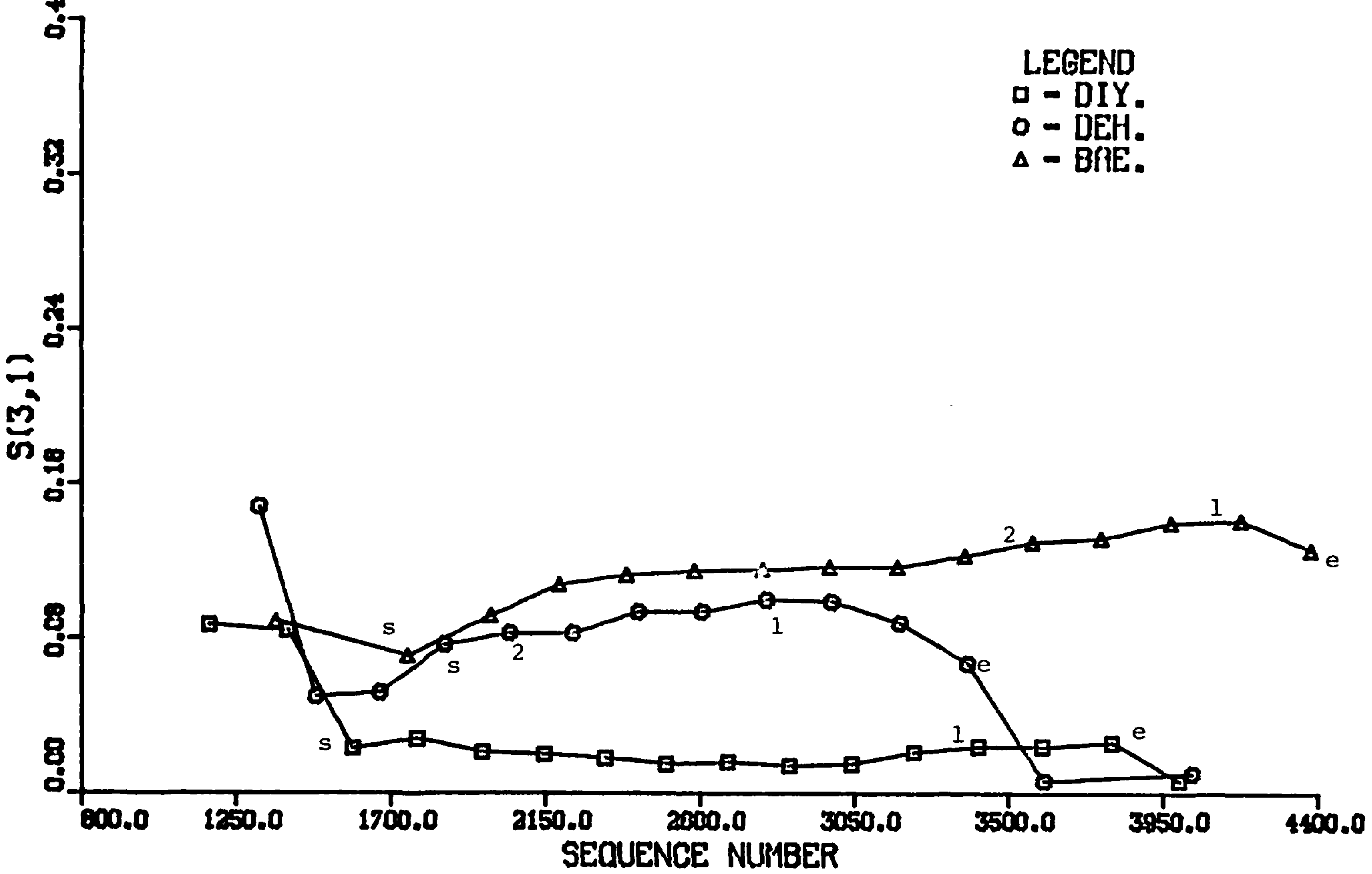
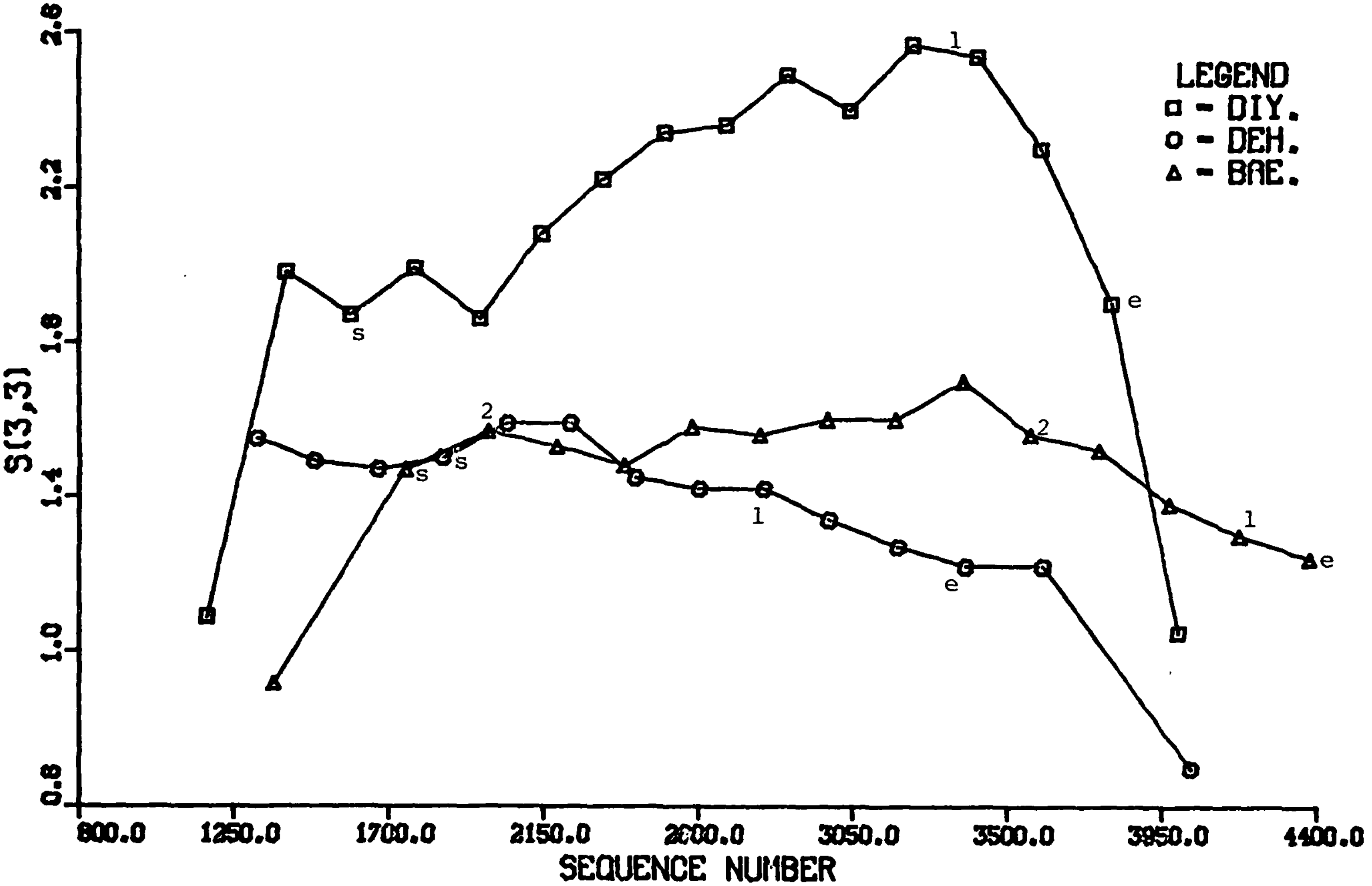

BACK VOWELS DUW,DAO,DAH (2236)
SEQUENCE NUMBER
CORNER RATIO 2
800.0
1250.0
1700.0
2150.0
2600.0
3050.0
3500.0
3950.0
4400.0
0.60
0.68
0.76
0.84
0.92
1.00
LEGEND
□ = DUW.
○ = DAO.
△ = DAH.
s
1
2
e

FIGURE 19.



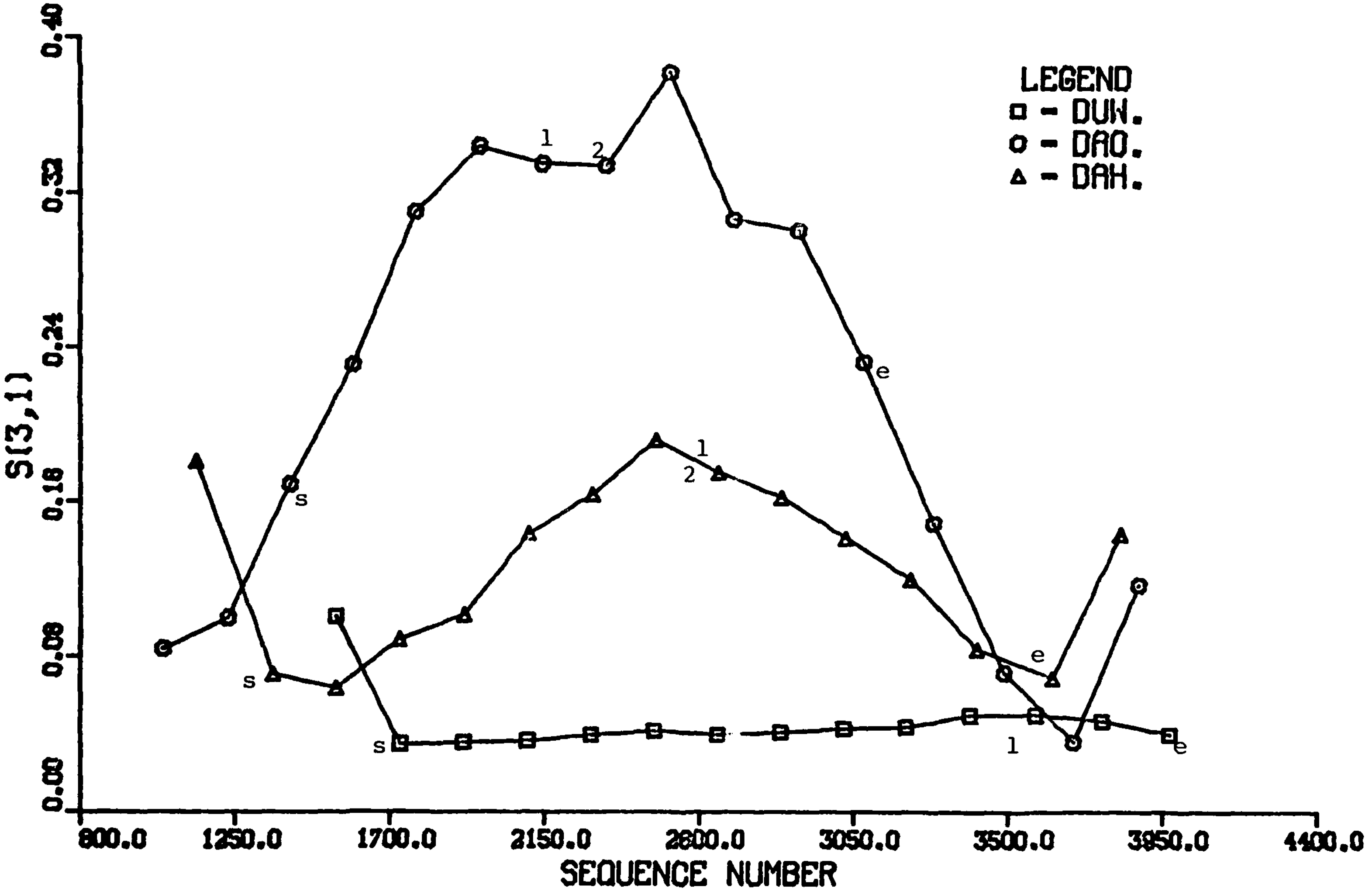
FRONT VOWELS DIY,DEH,BAE (236)

FIGURE 20.



FRONT VOWELS DIY,DEH,BAE (236)

FIGURE 21.



BACK VOWELS DUW,DAO,DAH (236)

FIGURE 22.

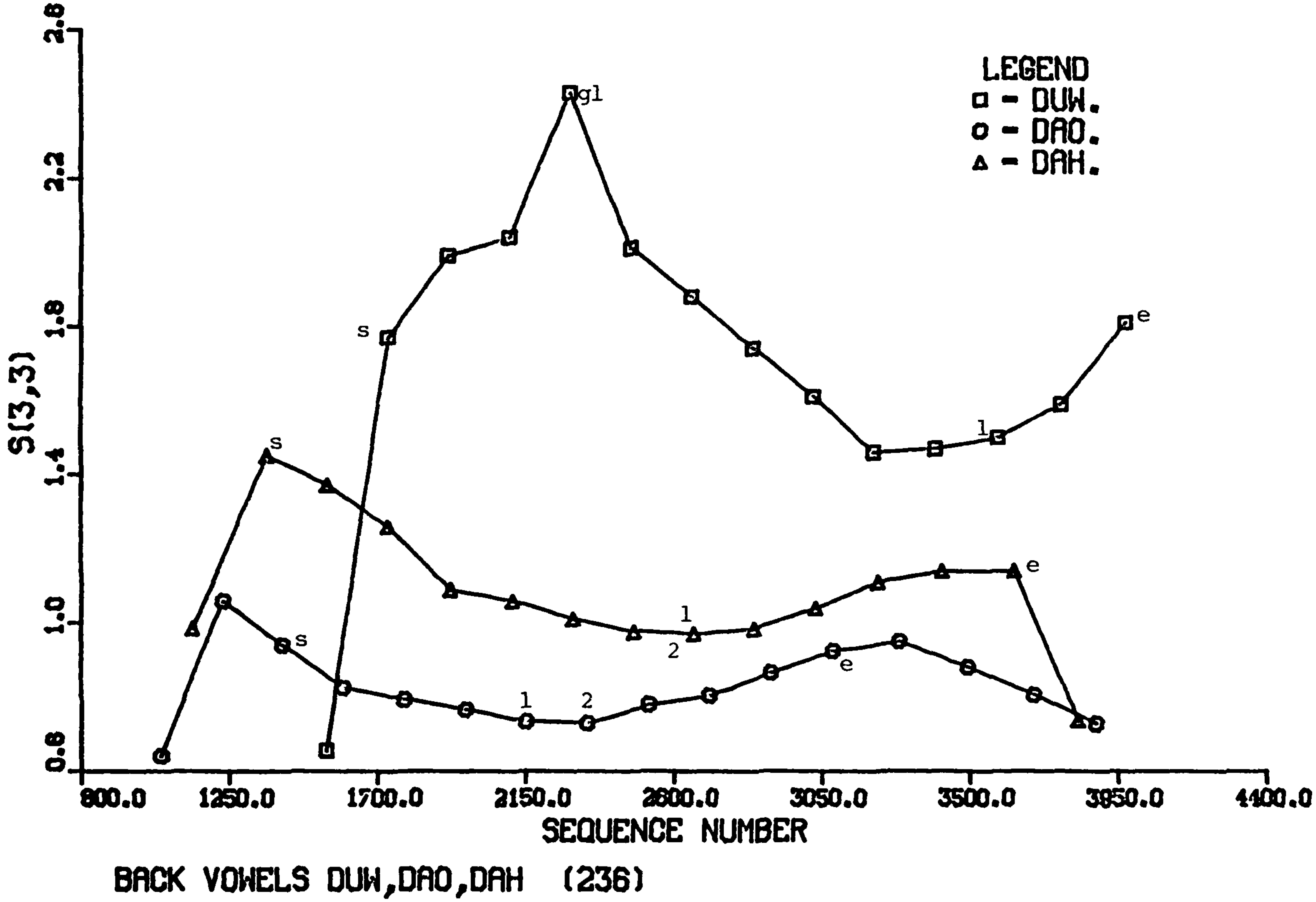
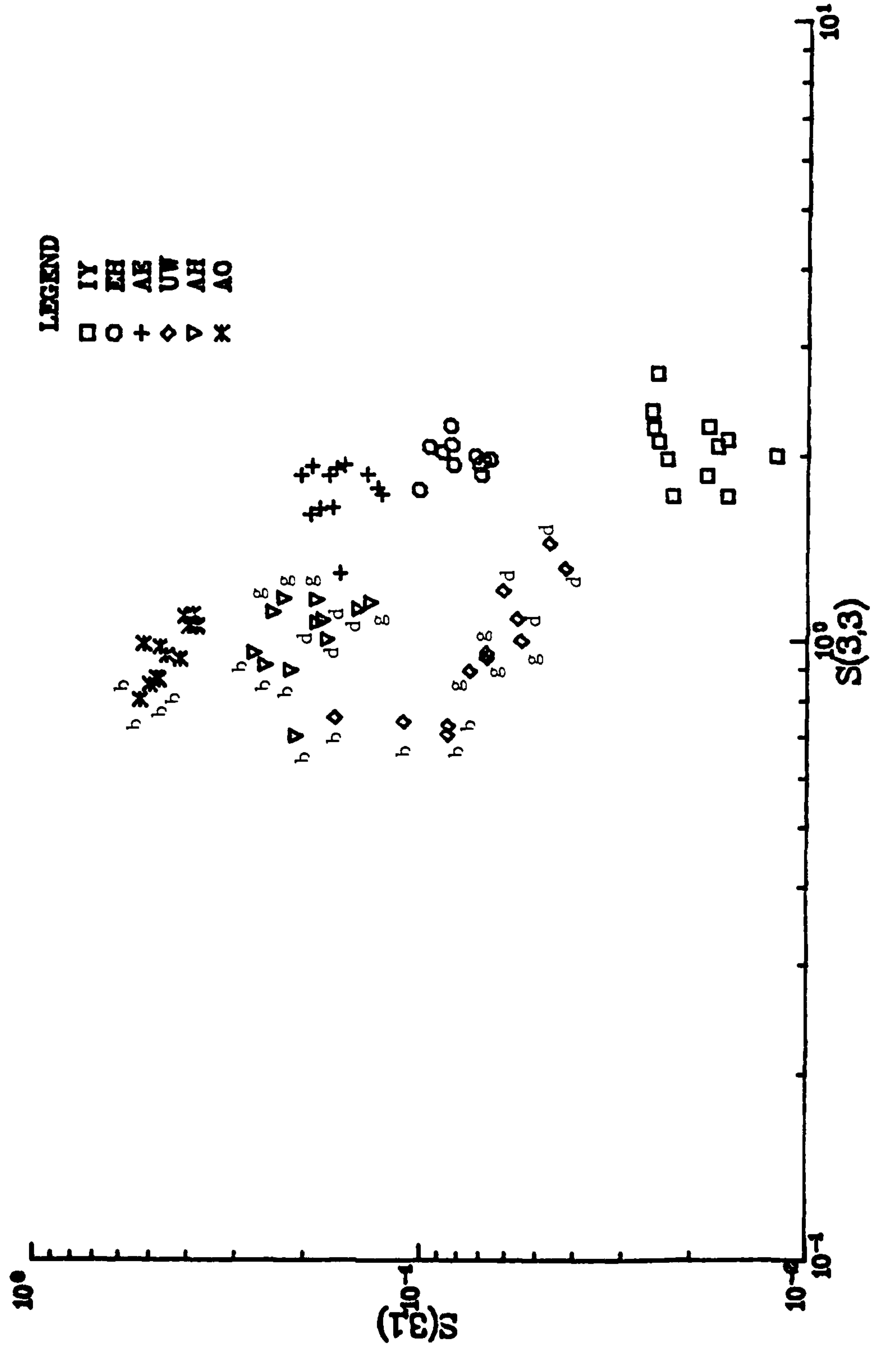

LEGEND
□ - DUW.
○ - DAO.
△ - DAH.
S(3,3)
0.6
1.0
1.4
1.8
2.2
2.6
800.0
1250.0
1700.0
2150.0
2600.0
3050.0
3500.0
3950.0
4400.0
SEQUENCE NUMBER
gl
s
e
1
s
e
1
2
s
e
1
2
BACK VOWELS DUW,DAO,DAH (236)

FIGURE 23.

Each vowel is indicated by a different symbol, and in no case is there an overlap among elements of the six vowel groups. These results indicate that it should be possible to accurately identify each of the phonemes.

Standard statistical methods have been used to analyze these data. The mean and standard deviation of the sensitivity elements S(3,1) and S(3,3) have been calculated for each vowel and for each consonant-vowel combination and are tabulated in Tables 3 through 8. Part of the variation within each vowel group was thought to reflect coarticulation effects. In order to detect and characterize any coarticulation, a multivariant analysis of variance was calculated for each consonant-vowel combination, and the corresponding P Value was also listed in the Tables 3 through 8. In Figures 24, 25, and 26, several groups of vowels with the same initial consonant are labeled to provide a graphical illustration of this coarticulation effect. Also, a summary of the significant consonant vowel coarticulation is given in Table 9. A systematic study of such coarticulation effects is proposed in the section titled Future Work.

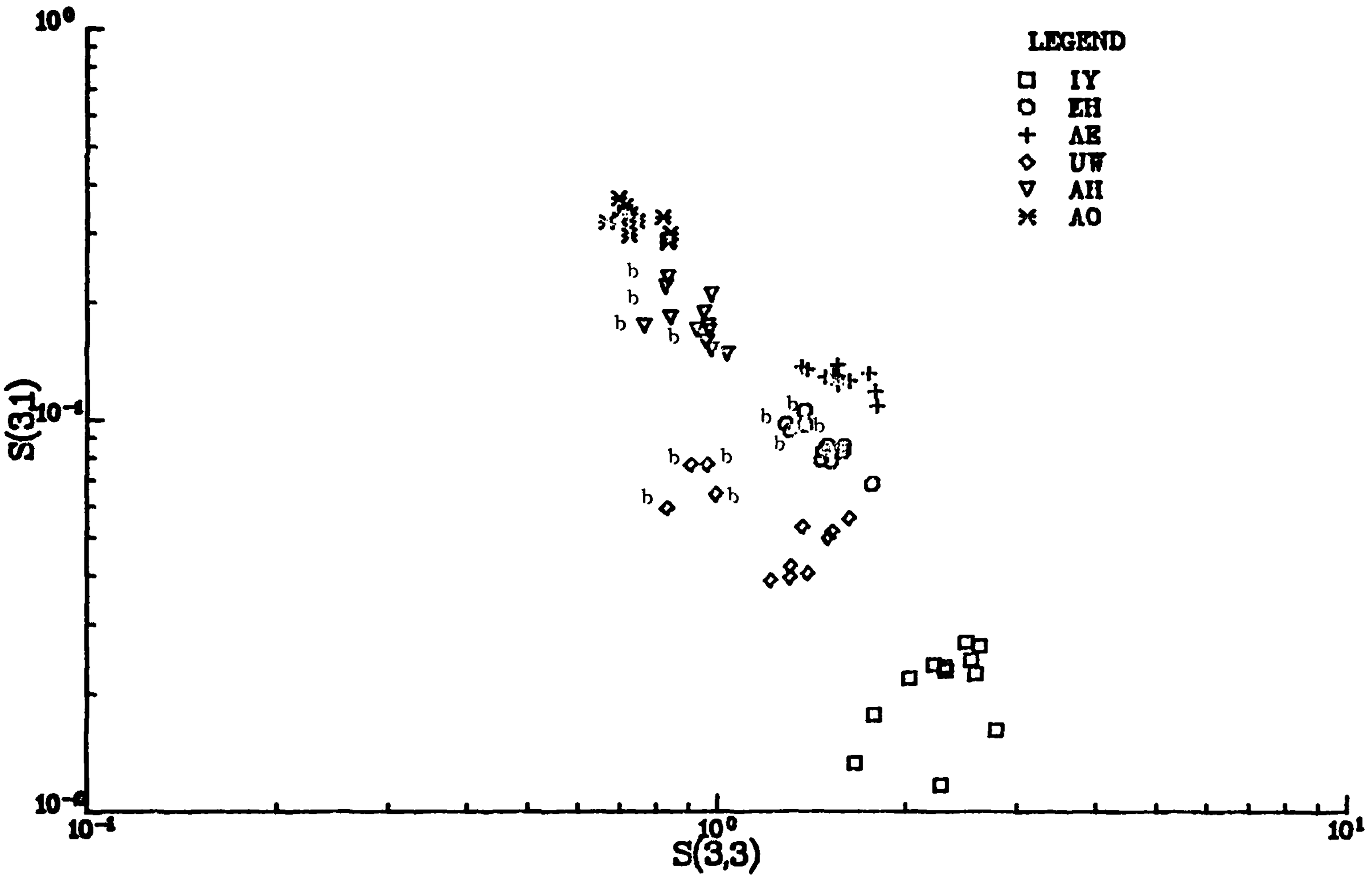
In order to assess the statistical differences between neighboring vowel pairs /EH, AE/, /AE, AH/, and /AH, AO/, a multivariant analysis of variance was calculated for each subject. The resulting P Values were listed in Table 10. In each case, at least one of the two sensitivity elements has a corresponding P Value of .0000. This result is consistent with the nonoverlapping nature of the sensitivity elements that comprise each vowel group. A summary of the minimum per-cent Euclidean separation between these vowel groups, in the S(3,1) vs. S(3,3) space, is shown in Table 11 and provides another description of their nonoverlapping nature.



SUBJECT #1

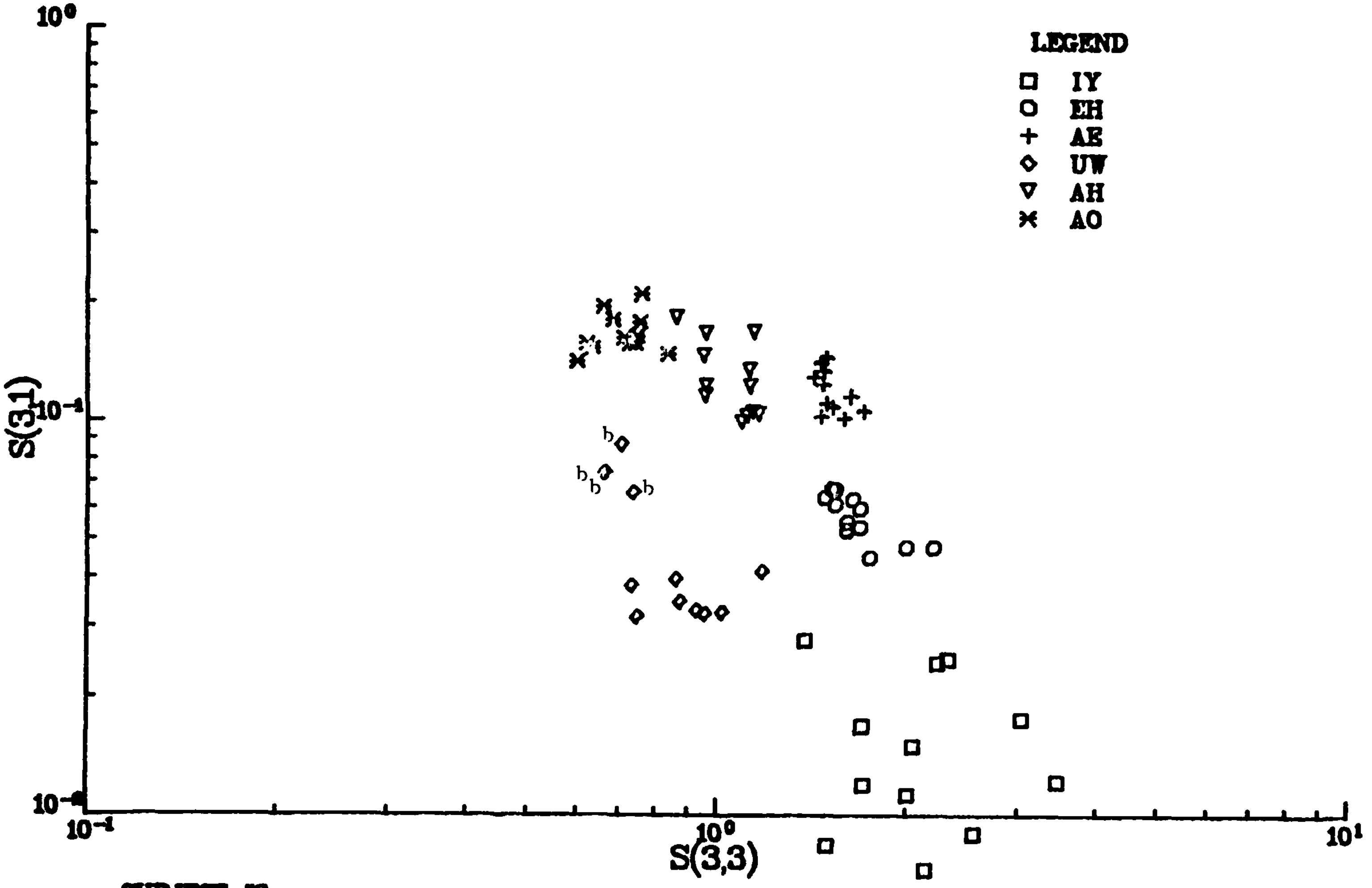
FRAME SELECTED VIA ALGORITHM 2

FIGURE 24.



SUBJECT #2
FRAME SELECTED VIA ALGORITHM 2

FIGURE 25.



SUBJECT #3

FRAME SELECTED VIA ALGORITHM 2

FIGURE 26.

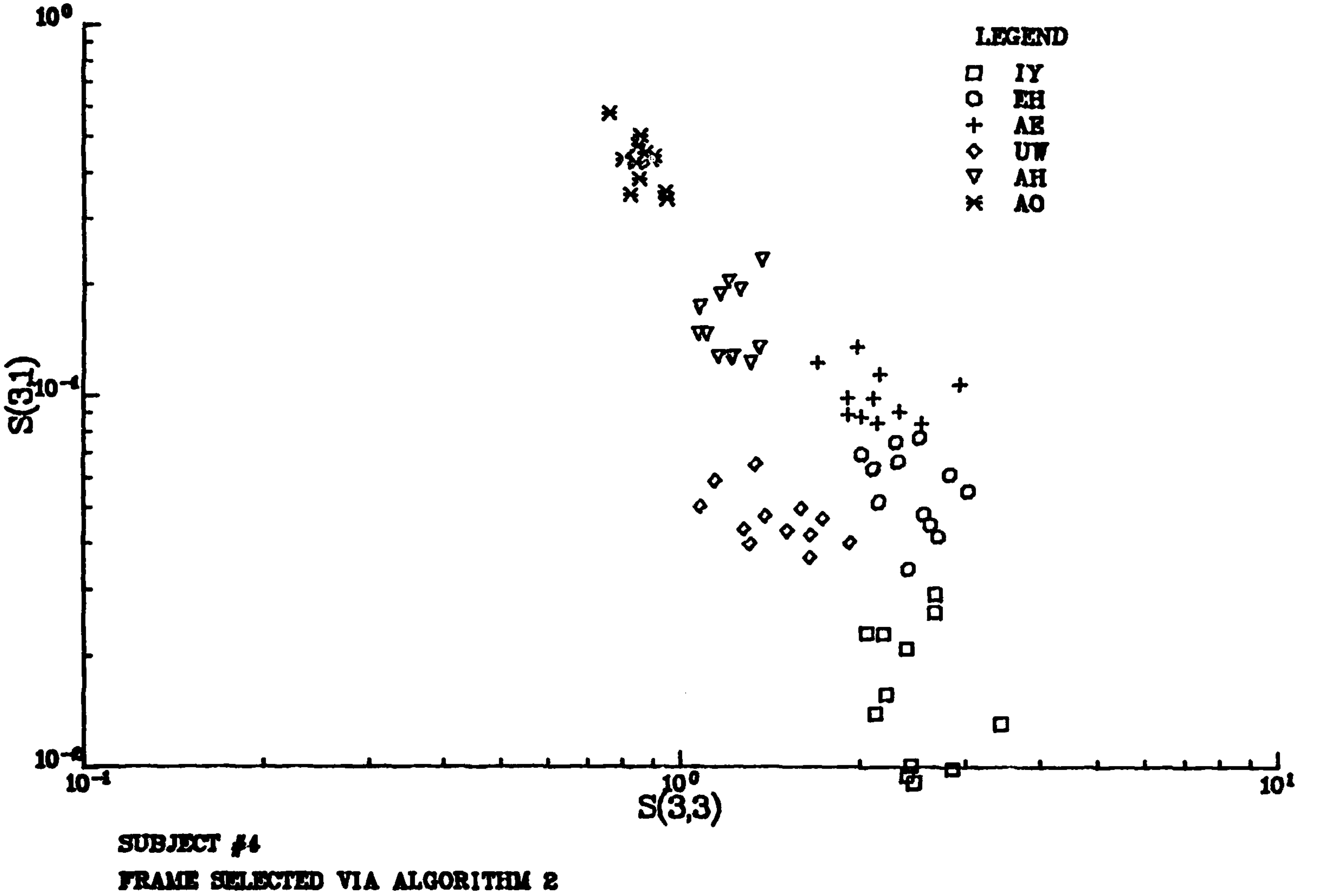
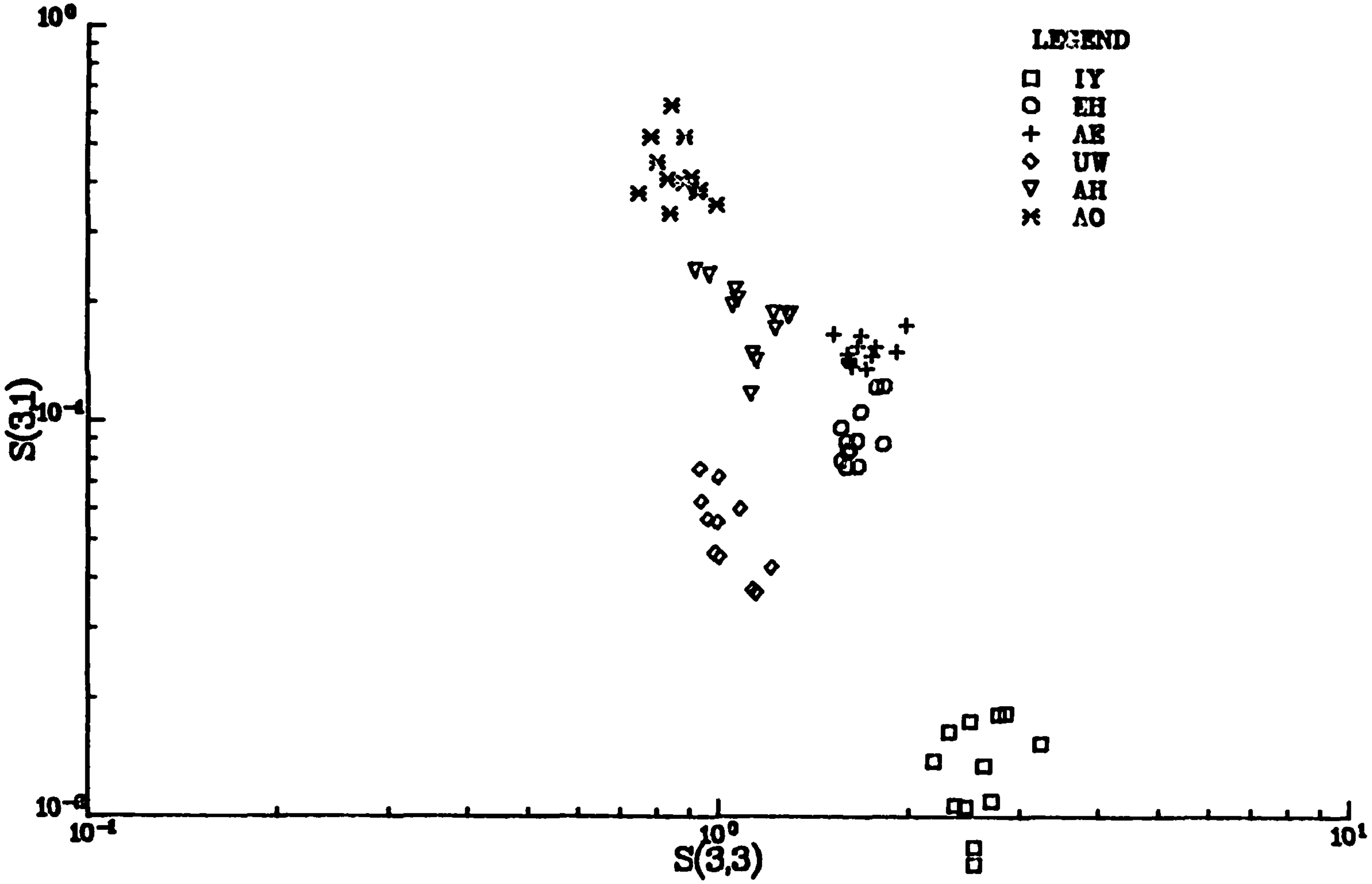

LEGEND
□ IY
○ EH
+ AE
◇ UW
▽ AH
✱ AO
10^0
10^{-1}
10^{-2}
S(3,1)
10^{-1}
10^0
10^1
S(3,3)
SUBJECT #4
FRAME SELECTED VIA ALGORITHM 2

FIGURE 27.



SUBJECT #5
FRAME SELECTED VIA ALGORITHM 2

FIGURE 28.

TABLE 3. MEAN AND STANDARD DEVIATION FOR VOWEL AND CONSONANT VOWEL COMBINATIONS AND MULTIVARIANT ANALYSIS OF VARIANCE FOR CONSONANT VOWEL COMBINATIONS

	SENSITIVITY	MEAN	(STANDARD DEVIATION)			
SUBJECT	ELEMENT	IY	BIY	DIY	GIY	P VALUE
1	S(3,1)	.0203	.0211	.0193	.0206	>.10
		(.0045)	(.0046)	(.0059)	(.0040)	
	S(3,3)	2.08	2.21	2.03	2.01	>.10
		(.28)	(.40)	(.26)	(.12)	
2	S(3,1)	.0210	.0225	.0217	.0189	>.10
		(.0051)	(.0046)	(.0067)	(.0043)	
	S(3,3)	2.30	2.47	2.41	2.01	>.10
		(.34)	(.22)	(.20	(.41)	
3	S(3,1)	.0156	.0185	.0192	.0090	.0467
		(.0069)	(.0041)	(.0083)	(.0017)	
	S(3,3)	2.19	2.27	2.23	2.06	>.10
		(.62)	(.57)	(.92)	(.44)	
4	S(3,1)	.0169	.0217	.0187	.0105	.0499
		(.0071)	(.0063)	(.0073)	(.0018)	
	S(3,3)	2.49	2.31	2.48	2.68	>.10
		(.38)	(.28)	(.33)	(.51)	
5	S(3,1)	.0134	.0145	.0107	.0151	>.10
		(.0038)	(.0045)	(.0024)	(.0032)	
	S(3,3)	2.60	2.52	2.58	2.68	>.10
		(.28)	(.27)	(.11)	(.44)	

TABLE 4. MEAN AND STANDARD DEVIATION FOR VOWEL AND CONSONANT VOWEL COMBINATIONS AND MULTIVARIANT ANALYSIS OF VARIANCE FOR CONSONANT VOWEL COMBINATIONS

	SENSITIVITY	MEAN	(STANDARD DEVIATION)			
SUBJECT	ELEMENT	EH	BEH	DEH	GEH	P VALUE
1	S(3,1)	.0814	.0828	.0764	.0852	>.10
		(.0115)	(.0156)	(.0134)	(.0026)	
	S(3,3)	1.97	1.91	1.96	2.06	>.10
		(.12)	(.11)	(.08)	(.13)	
2	S(3,1)	.0871	.0991	.0818	.0803	.0021
		(.0103)	(.0049)	(.0035)	(.0079)	
	S(3,3)	1.48	1.34	1.54	1.57	.0118
		(.13)	(.05)	(.06)	(.14)	
3	S(3,1)	.0572	.0649	.0582	.0484	.0003
		(.0078)	(.0027)	(.0044)	(.0033)	
	S(3,3)	1.70	1.53	1.67	1.90	.0241
		(.21)	(.03)	(.04)	(.27)	
4	S(3,1)	.0575	.0567	.0457	.0700	.0200
		(.0137)	(.0125)	(.0088)	(.0074)	
	S(3,3)	2.47	2.24	2.66	2.50	>.10
		(.31)	(.31)	(.27)	(.25)	
5	S(3,1)	.0933	.0913	.0809	.1078	.0388
		(.0161)	(.0103)	(.0059)	(.0180)	
	S(3,3)	1.66	1.62	1.62	1.75	.0827
		(.10)	(.04)	(.05)	(.12)	

TABLE 5. MEAN AND STANDARD DEVIATION FOR VOWEL AND CONSONANT VOWEL COMBINATIONS AND MULTIVARIANT ANALYSIS OF VARIANCE FOR CONSONANT VOWEL COMBINATIONS

	SENSITIVITY	MEAN (STANDARD DEVIATION)				
SUBJECT	ELEMENT	AE	BAE	DAE	GAE	P VALUE
1	S(3,1)	.1670	.1705	.1598	.1708	>.10
		(.0254)	(.0351)	(.0149)	(.0285)	
	S(3,3)	1.74	1.84	1.66	1.71	>.10
		(.18)	(.09)	(.28)	(.12)	
2	S(3,1)	.1283	.1334	.1233	.1282	>.10
		(.0082)	(.0047)	(.0120)	(.0039)	
	S(3,3)	1.58	1.50	1.63	1.62	>.10
		(.14)	(.08)	(.21)	(.09)	
3	S(3,1)	.1218	.1330	.1197	.1127	>.10
		(.0160)	(.0131)	(.0166)	(.0145)	
	S(3,3)	1.52	1.51	1.48	1.59	>.10
		(.08)	(.09)	(.01)	(.10)	
4	S(3,1)	.0997	.1051	.0899	.1042	>.10
		(.0168)	(.0169)	(.0060)	(.0227)	
	S(3,3)	2.16	1.98	2.04	2.45	.0729
		(.33)	(.22)	(.10)	(.40)	
5	S(3,1)	.1518	.1488	.1508	.1559	>.10
		(.0121)	(.0069)	(.0120)	(.0176)	
	S(3,3)	1.70	1.66	1.66	1.78	>.10
		(.13)	(.08)	(.17)	(.14)	

TABLE 6. MEAN AND STANDARD DEVIATION FOR VOWEL AND CONSONANT VOWEL COMBINATIONS AND MULTIVARIANT ANALYSIS OF VARIANCE FOR CONSONANT VOWEL COMBINATIONS

	SENSITIVITY	MEAN	(STANDARD DEVIATION)			
SUBJECT	ELEMENT	UW	BUW	DUW	GUW	P VALUE
1	S(3,1)	.0769	.1124	.0519	.0664	.0134
		(.0343)	(.0388)	(.0088)	(.0084)	
	S(3,3)	.976	.730	1.26	.943	.0001
		(.240)	(.019)	(.15)	(.044)	
2	S(3,1)	.0545	.0698	.0476	.0471	.0039
		(.0134)	(.0088)	(.0082)	(.0069)	
	S(3,3)	1.25	.926	1.43	1.38	.0004
		(.26)	(.072)	(.17)	(.10)	
3	S(3,3)	.0490	.0756	.0350	.0364	.0000
		(.0204)	(.0090)	(.0043)	(.0036)	
	S(3,3)	.841	.693	1.02	.806	.0009
		(.161)	(.037)	(.11)	(.075)	
4	S(3,1)	.0472	.0548	.0407	.0461	.0314
		(.0083)	(.0095)	(.0029)	(.0042)	
	S(3,3)	1.47	1.21	1.68	1.51	.0086
		(.25)	(.12)	(.17)	(.19)	
5	S(3,1)	.0526	.1679	.0387	.0512	.0002
		(.0136)	(.0076)	(.0029)	(.0059)	
	S(3,3)	1.04	.987	1.158	.984	.0006
		(.09)	(.070)	(.035)	(.018)	

TABLE 7. MEAN AND STANDARD DEVIATION FOR VOWEL AND CONSONANT VOWEL COMBINATIONS AND MULTIVARIANT ANALYSIS OF VARIANCE FOR CONSONANT VOWEL COMBINATIONS

	SENSITIVITY		MEAN (STANDARD DEVIATION)			
SUBJECT	ELEMENT	AH	BAH	DAH	GAH	P VALUE
1	S(3,1)	.203	.238	.173	.199	.0633
		(.042)	(.028)	(.018)	(.048)	
	S(3,3)	1.027	.865	1.067	1.15	.0011
		(.141)	(.113)	(.050)	(.026)	
2	S(3,1)	.182	.203	.163	.181	.0919
		(.027)	(.028)	(.012)	(.025)	
	S(3,3)	.924	.821	.988	.964	.0001
		(.083)	(.036)	(.037)	(.026)	
3	S(3,1)	.132	.132	.109	.156	.0675
		(.029)	(.037)	(.009)	(.017)	
	S(3,3)	1.05	.973	1.14	1.05	.0512
		(.10)	(.096)	(.021)	(.10)	
4	S(3,1)	.161	.164	.126	.192	.0219
		(.037)	(.021)	(.002)	(.041)	
	S(3,3)	1.21	1.11	1.23	1.30	.0059
		(.10)	(.04)	(.06)	(.08)	
5	S(3,1)	.187	.224	.150	.188	.0033
		(.037)	(.020)	(.029)	(.013)	
	S(3,3)	1.12	.998	1.17	1.20	.0135
		(.12)	(.069)	(.08)	(.09)	

TABLE 8. MEAN AND STANDARD DEVIATION FOR VOWEL AND CONSONANT VOWEL COMBINATIONS AND MULTIVARIANT ANALYSIS OF VARIANCE FOR CONSONANT VOWEL COMBINATIONS

	SENSITIVITY	MEAN	(STANDARD DEVIATION)			
SUBJECT	ELEMENT	AO	BAO	DAO	GAO	P VALUE
1	S(3,1)	.454	.500	.428	.434	.0907
		(.053)	(.025)	(.064)	(.035)	
	S(3,3)	.958	.844	1.015	1.014	.0047
		(.100)	(.031)	(.075)	(.068)	
2	S(3,1)	.322	.342	.313	.309	>.10
		(.026)	(.023)	(.026)	(.020)	
	S(3,3)	.755	.698	.782	.785	.0733
		(.063)	(.020)	(.070)	(.055)	
3	S(3,1)	.168	.173	.174	.158	>.10
		(.020)	(.020)	(.029)	(.007)	
	S(3,3)	.708	.652	.708	.765	.0478
		(.069)	(.028)	(.075)	(.050)	
4	S(3,1)	.431	.447	.441	.406	>.10
		(.069)	(.096)	(.049)	(.068)	
	S(3,3)	.865	.809	.880	.906	.0181
		(.055)	(.036)	(.027)	(.051)	
5	S(3,1)	.433	.428	.495	.377	>.10
		(.086)	(.078)	(.104)	(.029)	
	S(3,3)	.861	.828	.862	.892	>.10
		(.070)	(.110)	(.035)	(.045)	

TABLE 9. CONSONANT VOWEL COARTICULATION

SUBJECT	SENSITIVITY ELEMENT	NUMBER OF SUBJECTS WITH P VALUE <.10 IY	EH	AE	AH	AO	UW
All five	S(3,1)	2	4	0	5	5	1
	S(3,3)	0	3	1	5	5	4

TABLE 10. MULTIVARIANT ANALYSIS OF VARIANCE FOR VOWEL PAIRS

SUBJECT	SENSITIVITY ELEMENT	P VALUE FOR VOWEL PAIR AH-AE	AE-AH	AH-AO
1	S(3,1)	.0000	.0102	.0000
	S(3,3)	.0011	.0000	.0000
2	S(3,1)	.0000	.0000	.0000
	S(3,3)	.0000	.0000	.0000
3	S(3,1)	.0000	>.10	.0010
	S(3,3)	.0028	.0000	.0000
4	S(3,1)	.0000	.0000	.0000
	S(3,3)	.0126	.0000	.0000
5	S(3,1)	.0000	.0001	.0000
	S(3,3)	>.10	.0000	.0000

TABLE 11. MINIMUM PER-CENT EUCLIDEAN SEPARATION, IN SENSITIVITY SPACE, BETWEEN VOWEL GROUPS

SUBJECT	MINIMUM PER-CENT EUCLIDEAN SEPARATION EH-AE	AE-AH	AH-AO
1	25	9	42
2	23	25	22
3	52	22	9
4	9	22	70
5	11	19	38

Discussion.

This research program was focused on a selected set of six vowels contained in single words spoken in a simple carrier phrase by five males with differing dialects. The objectives were to evaluate the sensitivity matrix, interpret its changes during the production of the vowels, and to evaluate inter-speaker variations.

As described in the Methods section, each test word was divided into successive frames. For each frame, the covariance method was used to determine the coefficients of an 18 pole speech model. From the corresponding signal spectrum, the first three formant frequencies were located using a peak-picking algorithm. These general purpose signal processing methods need further study and development. In particular, there are occasional frames within the vowel portion of a word where the signal processing results seem inexplicable. An example of such a "glitch" is labeled gl in Figure 23.

Under the hypothesis of a lossless model, these first three formant frequencies were used to calculate Corner Ratio 1, Corner Ratio 2, and the sensitivity elements S(3,1) and S(3,3) for each frame. As defined by Equations 11 and 12, Corner Ratio 1 is large when formants 1 and 2 are close; whereas, Corner Ratio 2 is large when formants 2 and 3 are close. Comparing Corner Ratio 1 curves for the three front vowels in Figure 16 shows the changes tht occur during the vowel portion of a word and shows the progressive formant 1, formant 2 shifts that occur among the high /IY/ to low /AE/ front vowels. Figure 18 shows similar Corner Ratio 1 results for the group of back vowels. Comparing Corner Ratio 2 curves for the three front vowels in Figure 17 shows they are generally large, rise to a maximum, and then slowly fall. As a contrast, Figure 19 shows the corresponding Corner Ratio 2 curves for the back vowels. For this subject, there is a clear minimum with progressive shift from the high /UW/ to low /AO/ back vowels. The general character of these curves and their differences across vowel groups has proved to be important.

It was found that the vowel interval could be segmented by placing limits on the values of Corner Ratio 1 and Corner Ratio 2 and on their smoothness. It was also found that Corner Ratio 1 and Corner Ratio 2 could be used to determine when a vowel was "on target" by describing the location of the formants relative to the root-locus corners CR1 and CR2 shown in Figure 14. Algorithm 1 selected a "target" frame based on the behavior of Corner Ratio 1. For some speakers, the sensitivity elements of the selected frame may be used to accurately identify each vowel. However, nearest neighbor ambiguity may arise if a speaker's dialect does not clearly distinguish between phonemes, if a speaker has a greater tendency to coarticulate, or if a speaker tends to diphthongize speech productions.

The results presented in Figures 24 through 28 were obtained by a two-level identification scheme where the second level used a vowel-pair-specific form of Algorithm 2 that examined the behavior of Corner Ratio 2. For each of the speakers, all 6 vowel groups are nonoverlapping. Tables 10 and 11 demonstrate this separation in statistical and geometric terms. The minimum per-cent Euclidean separation varies across speakers as well as vowel-pairs. In the case of Subject 1, the 9 per-cent separation between AE and AH is due to a "single wild point." For speaker 5, the 11 per-cent separation between EH and AE appears to be due to coarticulation in words that begin with /g/. Subjects 3 and 4 have cases of 9 per-cent separation that appear to be due to their individual dialects. The EH-AE separation for speaker 4, from Virginia, appears to be reduced because of co-articulation and a tendency to diphthongize all productions. Speaker 3, from Rhode Island, had a small AH-AO separation because he used a consistent /ɒ/ production for /AO/.

The results presented in Figures 24 through 28 also provide a graphical view of the inter-speaker variations as well as the inter-speaker similarities. The differences in relative location of /AO/ for speakers 1 and 3 are consistent with the "live" judgment of the expert phoneticians; toward /AH/ for speaker 3 and /AA/ for speaker 1. These results reflect "phonetic facts of life" and indicate that accurate phoneme identification will require a "training" set of data for each speaker.

In summary, the specific objectives have been met. The sensitivity matrix was evaluated for the set of test words and speakers. It was found that Corner Ratio 1 and Corner Ratio 2 can be used to (a) segment the vowel interval and (b) locate when a vowel is "on target." It was also found that the sensitivity elements S(3,1) and S(3,3) of the "on target" frame should provide sufficient information to accurately identify each vowel within the test set.

VI. FUTURE WORK:

Any particular method of segmentation and identification of phonemes should be challenged by speech material which presents, in both a controlled and naturalistic manner, as many factors known to cause acoustic-phonetic variation as possible. A realistic expansion of the set of phonemes would include the unvoiced stop consonants /p, t, k/ and the nasal consonants /m,n/. The six stop consonants have the same manner, different voicing, differing place and should induce substantial coarticulatory variations. Likewise, the nasal consonants should induce substantial coarticulatory variations in the neighboring vowels.

The segmentation and identification studies would also be expanded to include both the consonants and the vowels. With this broader class of sounds, it is anticipated that it will be necessary to use improved spectral estimation techniques to obtain appropriate pole-zero models.[29-32] The root sensitivity analysis currently used for the formants can also be used for the zeros in the speech model.

The sequence of studies could be described as follows. Task I should investigate improved spectral estimation techniques for pole-zero models, evaluate the corresponding sensitivity matrix and the changes that occur during the production of the vowels in the words, test the conclusions reached in the current study, and evaluate inter-speaker variations. Task II should evaluate and interpret changes in the sensitivity matrix for the vowels due to coarticulation. Task III should evaluate and interpret the sensitivity matrix for the differing initial stop and nasal consonants. Finally, Task IV should use the above results to build a reference library, and should evaluate the efficacy of the sensitivity matrix in terms of the accuracy of the resulting phonetic representation of unknown speech.

VII. REFERENCES:

1. Air Force Systems Command Research Planning Guide (Research Objectives), HQ AFSC TR 82-01, pp. 7-24.

2. J. Shoup and L. Pfeifer, "Acoustic Characteristics of Speech Sounds," in J. Lass (ed.), Contemporary Issues in Experimental Phonetics (New York: Academic Press, 1976).

3. A. Liberman et al., "Perception of the Speech Code," Psych. Rev., 74 (1967), 431-61.

4. P. Liberman, Speech Physiology and Acoustic Phonetics (New York: Macmillan, 1972).

5. R. Daniloff and R. Hammerberg, "On Defining Coarticulation," J. of Phonetics, 1 (1972), 239-48.

6. K. Moll and R. Daniloff, "An Investigation of Velar Movements during Speech," J.A.S.A., 50 (1971), 678-84.

7. R. Kent, P. Carney and L. Severeid, "Velar Movement and Timing: Evaluation of a Model for Binary Control, J. Speech and Hearing Research, 17 (1974), 470-88.

8. A. Liberman, P. Delattre and F. Cooper, "The Role of Selected Stimulus Variables in the Perception of the Unvoiced Stop Consonants," American J. of Psychology, 65 (1952), 497-516.

9. P. Ladefoged, A Course in Phonetics (New York: Harcourt, Brace and Jovanovich, 1975).

10. R. Kent and R. Netsell, "Effects of Stress Contrasts on Certain Articulatory Parameters," Phonetica, 24 (1972), 23-44.

11. G. Fant, Acoustic Theory of Speech Production (The Hague: Mouton, 1970).

12. Dennis H. Klatt, "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access," J. of Phonetics (1979), pp. 1-34.

13. Dennis H. Klatt, "Review of the ARPA Speech Understanding Project," J. Acoust. Soc. Am., 62 (1977), 1345-66.

14. Dennis H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer," unpublished course notes from Speech Communication (Massachusetts Institute of Technology, June 18-23, 1979), p. 291.

15. R. Tomovic and M. Vukobratovic, General Sensitivity Theory (New York: American Elsevier, 1972).

16. Benjamin C. Kuo, Automatic Control Systems (4th ed., Englewood Cliffs, N.J.: Prentice-Hall, 1982), chap. 7.

17. Richard Absher, "Sensitivity Based Segmentation and Identification in Automatic Speech Recognition," (Final Report, 1982 USAF-SCEEE Summer Faculty Research Program, Contract No.: F49620-82-C-0035).

18. Kenneth N. Stevens, "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data," in P. B. Denes and E. E. David, Jr., Human Communication, A Unified View (New York: McGraw-Hill, 1972), pp. 51-66.

19. J. Makhoul, "Linear Prediction: A Tutorial Review," IEEE Proceedings, 63, No. 4 (April 1975), 561-80.

20. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am., 50 (1971), 637-55.

21. J. D. Markel, "Application of a Digital Inverse Filter for Automatic Formant and F_o Analysis," IEEE Trans. on Audio and Electroacoustics, AU-21, No. 3 (June 1973), 149-53.

22. J. D. Markel, "Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation," IEEE Trans. on Audio and Electroacoustics, AU-20, No. 2 (June 1972), 129-37.

23. S. S. McCandless, "An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra," IEEE Trans. on Acoustics, Speech, and Signal Proc., ASSP-22, No. 2 (April 1974), 135-41.

24. Programs for Digital Signal Processing, IEEE Press, (New York, 1979).

25. S. Chandra and W. C. Lin, "Experimental Comparison between Stationary and Non-Stationary Formulations of Linear Prediction Applied to Speech," IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-22 (1974), 403-15.

26. J. D. Markel and A. H. Gray, Jr., Linear Prediction of Speech, (New York: Springer-Verlag, 1976).

27. J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimates," IEEE Trans. on Audio and Electroacoustics, AU-20, No. 5 (December 1972), 367-77.

28. C. Thomas, Phonetics of American English (New York: Ronald Press Inc., 1958).

29. K. Steiglitz, "On the Simultaneous Estimation of Poles and Zeros in Speech Analysis," IEEE Trans. Acoustics, Speech, and Signal Proc., ASSP-25 (June 1979), 229-34.

30. L. R. Rabiner and R. W. Schafer, "Homomorphic Speech Processing," in Digital Processing of Speech Signals (Prentice-Hall, 1978), pp. 355-90.

31. Ira S. Konvalinka, R. Miroslav, and R. Matausek, "Simultaneous Estimation of Poles and Zeros in Speech Analysis and ITIF-Iterative Inverse Filtering Algorithms," IEEE Trans. on Acoustics, Speech, and Signal Proc., ASSP-27, No. 5 (October 1979), 485-92.

32. Ira S. Konvalinka, "Iterative Nonparametric Spectrum Estimation," IEEE Trans. on Acoustics, Speech, and Signal Proc., ASSP-32, No. 1 (February 1984), 59-69.

33. N. Rex Dixon and Thomas B. Martin (eds.), Automatic Speech & Speaker Recognition (Selected Reprint Series, New York: IEEE Press, 1979).

VIII. PERSONNEL:

The principal investigator, Dr. Richard Absher, devoted 23.5 percent effort to the research program. As his thesis research, one M.S. candidate, Thomas Saunders, worked extensively on the research program. His thesis will be entitled "A Sensitivity Measure for a Voice Wave Analysis."

Dr. Raymond Daniloff, Professor and Chairman, and Dr. Paul R. Hoffman, Assistant Professor, both of the Department of Communication Science and Disorders, have served as expert phoneticians during the recording sessions and have provided guidance on acoustic-phonetic problems.

IX. PUBLICATIONS:

Based on the results described in this report, a paper will be submitted for publication in the Journal of the Acoustical Society of America. The proposed authors and title are: Richard G. Absher and Thomas Saunders, "Sensitivity Based Segmentation and Identification of Vowels in Continuous Speech."